

Article

Clustering Approach for the Efficient Solution of Multiscale Stochastic Programming Problems: Application to Energy Hub Design and Operation under Uncertainty

Mohammed Alkatheri ¹, Falah Alhameli ¹, Alberto Betancourt-Torcat ¹, Ali Almansoori ^{2,*}  and Ali Elkamel ^{1,2,*} 

¹ Department of Chemical Engineering, University of Waterloo, 200 University Ave. West, Waterloo, ON N2L 3G1, Canada

² Department of Chemical Engineering, Khalifa University of Science and Technology, Sas Al Nakhl Campus, Abu Dhabi P.O. Box 2533, United Arab Emirates

* Correspondence: ali.almansoori@ku.ac.ae (A.A.); aelkamel@uwaterloo.ca (A.E.); Tel.: +1-519-888-4567-37157 (A.E.)

Abstract: The management of the supply chain for enterprise-wide operations generally consists of strategic, tactical, and operational decision stages dependent on one another and affecting various time scales. Their integration usually leads to multiscale models that are computationally intractable. The design and operation of energy hubs faces similar challenges. Renewable energies are challenging to model due to the high level of intermittency and uncertainty. The multiscale (i.e., planning and scheduling) energy hub systems that incorporate renewable energy resources become more challenging to model due to an integration of the multiscale and high level of intermittency associated with renewable energy. In this work, a mixed-integer programming (MILP) superstructure is proposed for clustering shape-based time series data featuring multiple attributes using a multi-objective optimization approach. Additionally, a data-driven statistical method is used to represent the intermittent behavior of uncertain renewable energy data. According to these methods, the design and operation of an energy hub with hydrogen storage was reformulated following a two-stage stochastic modeling technique. The main outcomes of this study are formulating a stochastic energy hub optimization model which comprehensively considers the design and operation planning, energy storage system, and uncertainties of DRERs, and proposing an efficient size reduction approach for large-sized multiple attributes demand data. The case study results show that normal clustering is closer to the optimal case (full scale model) compared with sequence clustering. In addition, there is an improvement in the objective function value using the stochastic approach instead of the deterministic. The present clustering algorithm features many unique characteristics that gives it advantages over other clustering approach and the straightforward statistical approach used to represent intermittent energy, and it can be easily incorporated into various distributed energy systems.

Keywords: clustering algorithm; multiscale; supply chain; computational complexity; energy hub



Citation: Alkatheri, M.; Alhameli, F.; Betancourt-Torcat, A.; Almansoori, A.; Elkamel, A. Clustering Approach for the Efficient Solution of Multiscale Stochastic Programming Problems: Application to Energy Hub Design and Operation under Uncertainty. *Processes* **2023**, *11*, 1046. <https://doi.org/10.3390/pr11041046>

Academic Editor: Jie Zhang

Received: 16 January 2023

Revised: 16 March 2023

Accepted: 22 March 2023

Published: 30 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Time is organized in terms of years, months, days, and hours. Each time term is distinguished by scale and combined to form the multiscale phenomena that is part of our daily life. This is an outcome of multiscale dynamics in the solar system [1]. Conventionally, most modeling approaches focus on a mono-scale perspective. When the macroscale behavior of a system is the focal point, the microscale is modeled applying constitutive interactions. On the other hand, if the microscale is the subject matter, one assumes that not a compelling thing occurs at a macroscale level. Accordingly, multiscale modeling helps to manage the restrictions of both methods (macro- and microscale) by targeting the efficiency of macroscale modeling while preserving the precision of microscale modeling

altogether. However, the simultaneous use of several scales and preceptive levels leads to more complex modeling approaches [1].

The integration of a supply chain's decision level is extremely important for improving investment returns. Planning and scheduling are generally performed separately even though they are interdependent. The integration of planning (e.g., design) and scheduling (e.g., operation) improves the decision-level management resulting in lower net costs. Yet, large-scale problems are formed as a result of different time scales integration that are typically computationally intractable. In order to address this problem, various modeling and solution approaches have been proposed (e.g., Yilmaz et al. 2019 [2]). The majority are problem-specific and are only valid for short-time timeframes. Accordingly, clustering arises as an effective and appropriate approach to deal with this type of problem by means of grouping similar inputs such as supply, demand, and price together. Input parameters typically are made up of multiple attributes such as the concurrent electricity and heat demand. This approach can significantly reduce the model size combined with the improved computational performance while maintaining solution accuracy.

The task of clustering is to discover the structure in sets of unlabeled data by means of grouping it into uniform groups. Particularly, the similarity of the within-group-object (i.e., within one cluster) is minimized while the between-group-object dissimilarity is maximized (i.e., between different clusters). An effective clustering can be measured by a high similarity/uniformity within group objects as well as a high dissimilarity/heterogeneity amongst groups. Objects (i.e., events, measurements) are commonly characterized by elements or vectors in a multidimensional space where every dimension identifies a specific quantifiable attribute (variable) that describes the object. Thus, an arranged set of objects could be represented conceptually as an $m \times n$ matrix. The symbols m and n denote the rows (one per object) and columns (one per attribute), respectively.

Energy hub models can be applied to different spatial scales, (e.g., from a single building to a large geographic region) as well as time scales. Particularly, energy hub modeling could be applied to different time scales from long-term planning (e.g., designing and sizing the energy conversion and storage units) to medium- or short-term planning (scheduling and operation).

Many studies have addressed the optimal scheduling and planning of energy hub systems. For example, [3] studied the daily scheduling of an energy hub including different generation and storage technologies. Ma et al. [4], proposed a deterministic MILP model aiming to minimize the daily operating costs (including electricity/gas and carbon emissions costs) of an energy hub. Lu et al. [5],) proposed a multi-objective optimization framework for the optimal operation of energy hub components considering uncertain households' behaviors. Taqvi et al. [6] carried out a study to determine the rooftop renewable energy potential for the optimal designing of an EV charging infrastructure using a multi-energy hub approach, but this study did not consider energy storage and a multiscale approach for size reduction. A recent study [7] investigated the effect of the demand response and the impact of carbon trading on a multi-objective optimization model that considered the operation cost, energy utilization efficiency, and consumption rate of renewable energy.

In addition, energy hub network operation has been explored in several studies; however, few studies have carried out the design and operation of urban energy systems based on the energy hub concept. Koltsaklis et al. in 2014 [8] studied the optimal design and operation of distributed energy systems (DEs), but the study did not consider renewable energy technologies. Maroufmashat, Sattari, et al. [9], proposed a deterministic model for the design and operation of DEs in urban areas including renewable energy sources. Economic and environmental considerations were investigated, but renewable resource uncertainties were not addressed. Mukherjee et al. [10], proposed a stochastic programming approach for the planning and operation of a power-to-gas energy hub. The study focused on assessing the benefits of power-to-gas energy storage while accounting for the uncertainty of fuel cell vehicles, refueling hydrogen, and the electricity price. Kotzur et al. [11]

employed time series aggregation (based on clustering algorithms) to the cluster demand, wind speed, and solar irradiance. The study shows that the application of clustering methods in energy hub optimization significantly reduces the model complexity. A mixed-integer linear programming model for offshore energy hubs was developed by [12]. They used different aggregation, clustering, and time sampling to address the multi-timescale aspects without reducing the actual demand data, however, the uncertainty of offshore wind was not considered. In 2023, Amry et al. [13], developed a strategy to carry out the optimal sizing for an EV workplace charging station considering PV and flywheel energy storage systems; nevertheless, the work did not consider the uncertainty associated with solar energy.

The aforementioned studies did not consider uncertainty in the distributed renewable energy sources (DRESs) which could lead to inaccurate decisions. Zhang et al. [14], developed a two-stage stochastic model for the optimal design and operation of combined cooling, heat, and power (CCHP) units. The study considered the uncertainties of loads and solar irradiation in different seasons, but it did not include energy storage systems (ESS). f et al. [15] developed a stochastic model for the operation and scheduling of energy hubs considering the DRES's uncertainties. The stochastic scenarios were generated using a Monte Carlo simulation and were reduced using the k-means clustering algorithm. However, only three scenarios per uncertain feature were considered in the optimization.

1.1. Research Gap and Contribution

The current literature covers several aspects of energy hub modeling such as the design, operation, greenhouse gas (GHG) emissions minimization, DRESs, ESS, renewable energy uncertainty, and demand size reduction. However, no studies have combined all features into a single model formulation. Accordingly, there is a knowledge gap that demands the development of a comprehensive stochastic optimization model that considers design and operation planning, ESS, and the DRES's uncertainties. In addition, there is a need to apply an efficient reduction size to large-sized multi-attribute demand data used as an input to stochastic energy hub modeling. In other words, the main novelty of this work is to combine all the research aspects into one model and present a general and practical framework to solve different types of multiscale energy systems with multiple attributes. Table 1 summarizes the research aspect of previous energy hubs, the research gap, and the current research contribution.

Table 1. Literature review summary on energy hub optimization problems, the research gap, and the current research contribution.

Study	Year	Research Aspects						
		Optimal Design	Optimal Operation	(GHG) Emission Saving	DRERs	ESSs	Uncertainty of Renewable Energy	Demand size Reduction
[3]	2016	✗	✓	✗	✗	✓	✗	✗
[16]	2017	✗	✓	✓	✓	✓	✗	✗
[7]	2023	✗	✓	✓	✓	✓	✗	✗
[17]	2018	✗	✓	✓	✓	✓	✗	✗
[4]	2017	✗	✓	✓	✓	✓	✗	✗
[5]	2020	✗	✓	✗	✓	✓	✗	✗
[8]	2014	✓	✓	✗	✗	✓	✗	✗
[18]	2016	✗	✓	✓	✓	✓	✗	✗
[9]	2016	✓	✓	✓	✓	✓	✗	✗
[19]	2017	✓	✓	✓	✓	✓	✗	✗
[10]	2017	✓	✓	✗	✗	✓	✗	✗
[11]	2018	✓	✓	✗	✓	✓	✗	✓
[20]	2017	✗	✓	✗	✓	✓	✓	✗
[14]	2019	✓	✓	✓	✓	✗	✓	✗
[15]	2020	✗	✓	✗	✓	✓	✓	✗
[12]	2022	✓	✓	✓	✓	✓	✗	✗
[13]	2023	✓	✓	✗	✓	✓	✗	✗
Current work	2022	✓	✓	✓	✓	✓	✓	✓

Table 1. Cont.

Study	Year	Research Aspects						
		Optimal Design	Optimal Operation	(GHG) Emission Saving	DRERs	ESSs	Uncertainty of Renewable Energy	Demand size Reduction
Research gap		Combined all research aspects of the analysis into one model.						
Current work contribution		<ul style="list-style-type: none"> • Developing a stochastic optimization which comprehensively considers the design and operation planning, energy storage systems, and uncertainties of DRERs; • Applying an efficient size-reduction approach to large-sized multiple attributes demand data which can be used as an input to the stochastic energy hub model. 						

1.2. Research Problem

Typically, planning and scheduling are both performed separately even though they are interdependent. However, the integration between these different time scales is the key to improving the efficiency and profit margins, as the integration of planning (e.g., design) and scheduling (i.e., operation) improves decision level management which results in lower net costs. However, the computational tractability arising from this integration makes it difficult to solve. For example, a very large and intractable problem will be formed if different time scales of the multiscale energy hub model are converted to the shortest planning period (i.e., detailed scheduling over a long duration). While relaxing some constraints, employing surrogate models, or using an averaging method, these might lead to infeasible operations (i.e., since detailed schedules cannot be obtained to meet the planned production targets) or an inaccurate system design [11,21].

An energy hub modeling approach can establish controllable and flexible energy due to the ability to integrate different types of DRESs and energy storage system (ESSs) [22]. Using DRESs in the energy sector can alleviate the impact of greenhouse gas emissions from other sectors (e.g., agriculture, industry, transportation) and play a key role in pathways towards deep decarbonization [23]. However, due to the uncertainty and variability of DRESs (e.g., solar photovoltaic and wind turbine), the advantages of energy hub systems to supply flexible power could be limited and diminished [24]. Therefore, modeling the energy hub by considering the uncertainties associated with these sources is crucial.

1.3. Research Motivations and Goals

This study attempts to address the following challenges (multiscale decision making, uncertainty and variability in DRERs) associated with energy hubs by:

1. Applying unique general mathematical programming-based clustering methods to reduce the multiple attribute demand data size that has the ability to attain normal and sequence clustering, change the internal clustering measure, and tune attribute weights.
2. Proposing a statistical method that models the uncertain behavior of renewable energy sources.
3. Formulating the energy hub system as a two-stage stochastic optimization.

The first goal of the present work was to overcome the problem associated with integrating different scales of an energy hub model by adopting a generic clustering approach. The goal of the clustering approach was to represent the days in a year that exhibit a similar trajectory with a reduced-sized typical day candidate (i.e., representative) of the operating year. A sufficient number of representative curves mean the representatives are able to provide a close enough solution to the full size (high in accuracy) model while also maintaining solution tractability.

The second goal of this work was to develop a two-stage stochastic optimization model for the design and operation of an energy hub system with hydrogen storage. A hydrogen storage system was selected due to its flexibility in offering different energy recovery pathways. For instance, hydrogen can be used to produce electricity through a fuel-cell, supplied for the hydrogen demand to a hydrogen vehicle, injected and distributed into the existing natural gas infrastructure. Two case studies are considered to optimally design

and operate the energy hub model, one without restriction on green-house gas (GHG) emissions, and another restricting the GHG emissions. A Weibull distribution statistical method was implemented to generate stochastic wind speed scenarios from wind speed data. To test the clustering efficiency, the cluster results were applied to the developed energy hub model and compared with the energy hub when the whole set of data was used. At the end of this paper, the efficiency of the stochastic approach was assessed.

2. Methods: Clustering Algorithms and Stochastic Scenario Generation

The first aim of this work was tackling the problem arising from the integration of different supply chain decision levels by developing a generic clustering method.

The goal was to decrease the model size by means of switching the annual days with regular days characteristic of the operational year. Even though clustering has been extensively used in several applications, demand patterns clustering has been weakly analyzed. For instance, demand patterns are particularly complex due to their multidimensional nature comprising shape (e.g., hourly demand curves trajectory), while time regularly exhibits diverse attributes (e.g., the energy hub's parallel heat and electricity demand). Figure 1 shows a conceptual schematic of the proposed clustering approach application to the multiscale decision-making problem. This analysis is based on a mathematical programming approach. For instance, the clustering algorithm for multidimensional attributes is formulated by applying mixed-integer programming (MIP) methods. Due to the complexity of the MILP clustering model, a heuristic size reduction derived from the full-scale MILP clustering approach was proposed to tackle computational issues.

Multiple attribute full size data (yearly days)

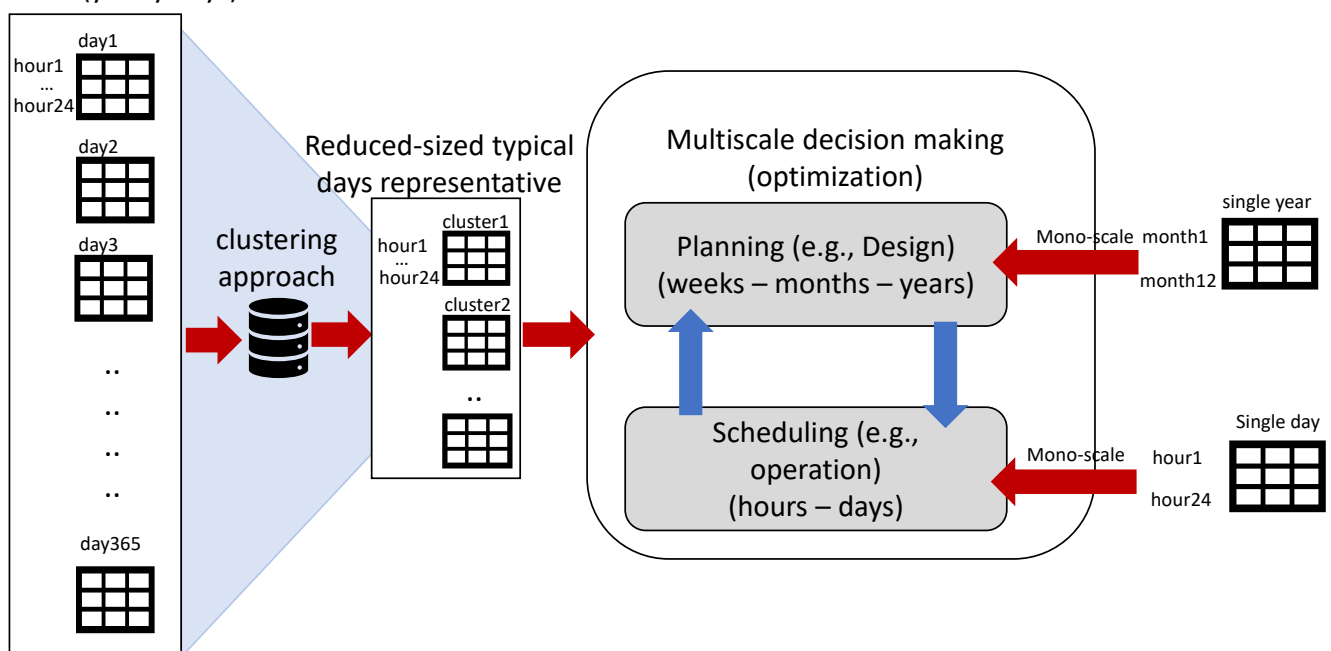


Figure 1. Clustering approach application to the multiscale decision-making problem.

The time-series data comprised the clustering algorithm proposed in the present work. Clustering has drawn significant attention given its potential processing applications for big data. The algorithm clusters demand data considered the trajectories-time and shape-similarity at once. Hence, the time-series data clustering could help with reducing the processing difficulties of multiscale modeling. For instance, the least absolute value method L_1 -norm [25–29] was applied to quantify the similarity and preserve the model's linearity while showcasing the generality of the algorithm.

The proposed clustering approach was an extension of the work previously proposed by [21] which included the clustering of multi-attributes as an alternative to the traditional single attribute. The weighting method has been selected as a multi-objective optimization approach [30] to cope with the problem's multiple-attribute nature (see Figure 2 that illustrates a bi-objective problem's Pareto front).

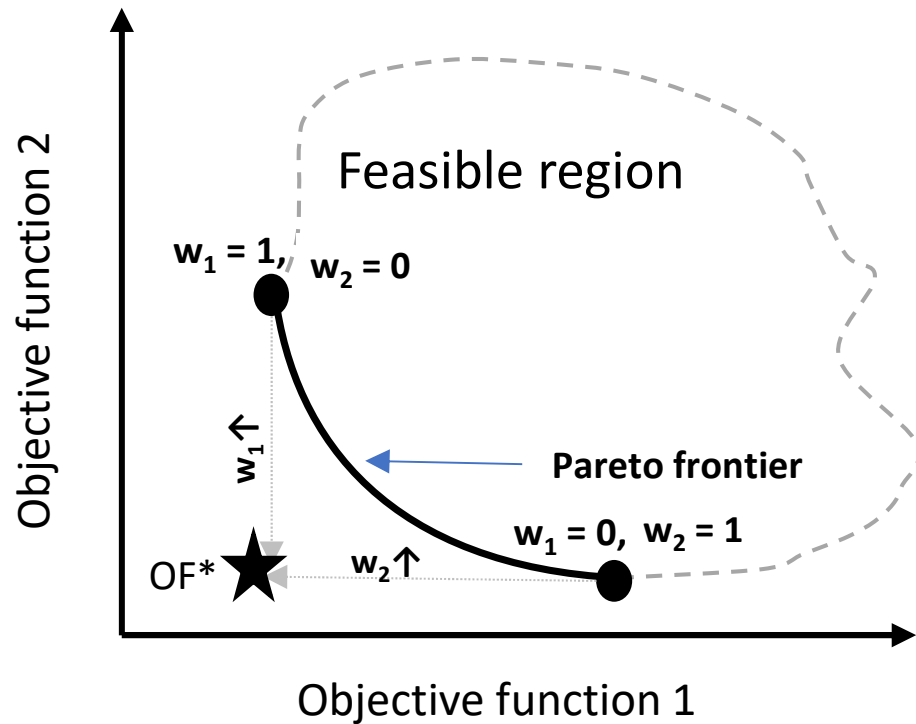


Figure 2. Illustration of Pareto frontier. The utopia point (OF^*) corresponds to the optimum of both objective functions 1 and 2. However, there is typically no feasible solution at utopia point as demonstrated in the figure.

2.1. General Algorithm Formulation

The clustering approach aims at allocating days to clusters with the least dissimilarity. Accordingly, the load curves set given in D (days) and H (hours) were collected in C clusters. Multiple attributes, denoted by index a , were considered within the model formulation. This can be expressed in the following general form:

$$\min Z = \sum_a w_a IAE_a \quad (1)$$

$$\text{s.t. } \sum_{c=1}^C x_{d,c} = 1, \quad \forall d \quad (2)$$

where Z is the multi-objective performance criteria function to be minimized, and IAE_a (Integral Absolute Error) denotes the L_1 -norm employed to measure similarity in the a^{th} attribute. Equation (1) shows the multi-objective performance criteria of different attributes a as a weighted function, with w_a as attribute a 's weighting factor ($w_a \geq 0, \sum_a w_a = 1$). In contrast, Equation (2) represents the constraint for the day assignment where every single day of the year must be allocated to a curves c cluster. The binary variable $x_{d,c}$ indicates the binary variable allocating loads for day d joining cluster c . The IAE formulation can be expressed as follows:

$$IAE = \int_a^b |L(t) - C(t)| dt \quad (3)$$

where $L(t)$ and $C(t)$ represent the load and clustered curves, respectively. Furthermore, Equation (4) results from employing the trapezoidal rule to Equation (3) as follows:

$$IAE_a = \frac{\Delta}{2} * \sum_{d=1}^D \sum_{h=1}^{H-1} AD_{a,d,h} + AD_{a,d,h+1}, \quad \forall a \quad (4)$$

where $AD_{a,d,h}$ denotes the absolute difference between the clustered curve c and load curve l for the h^{th} hour of d day for attribute a , which can be defined as follows:

$$AD_{a,d,h} \geq |DL_{a,d,h} - D_{a,c,h}| \times x_{d,c}, \quad \forall a, h, d, c, DL_{a,d,h} \quad (5)$$

$DL_{a,d,h}$ denotes a 's attribute demand load for hour h in the d^{th} day, $D_{a,c,h}$ is the representative demand of attribute a for the h^{th} hour in the c cluster. It is worth noticing that the model can be adapted to different performance criteria. For instance, the L_2 norm can be easily implemented in place of the L_1 norm by incorporating the Euclidean distance shown in Equation (3).

Additionally, the demand data could be sequentially clustered by incorporating a constraints set according to the string property concept [31]. Sequence clustering could assist with maintaining flexible operations, such as for cases arising where similar continuous operations are preferred to minimize the change-mode or set ups costs. Consequently, the constraints sets are included to integrate the time-dimension sequencing into the clustering model:

$$x_{d+1,1} \leq x_{d,1}, \quad \forall d < D \quad (6)$$

$$x_{d+1,c} \leq x_{d,c} + x_{d,c-1}, \quad \forall d < D, c > 1 \quad (7)$$

$$x_{D,c} \leq x_{D-1,c} + x_{D-1,c-1}, \quad \forall c > 1 \quad (8)$$

Equations (7)–(9) control the initial, intermediate, and final sequence clusters, respectively.

The proposed general formulation provides the same platform to perform sequence and normal clustering because it is built on the same algorithmic structure. Nevertheless, the model is an MINLP given the multiplication of variable $D_{a,c,h}$ and $x_{d,c}$ (shown in Equation (5)) and the absolute value. Hence, the absolute function is linearized by applying linearization methods on the absolute function [32]. Furthermore, the bilinear term ($D_{a,c,h} x_{d,c}$) is further linearized by incorporating an extra continuous variable ($RV_{a,h,d,c} = D_{a,c,h} * x_{d,c}$) called the relaxation variable, through a set of constraints [33]. Further details on the linearization approach can be found in [21]. In summary, the model for normal clustering is made up by (1)–(3), whereas sequence clustering is denoted by (1)–(3), and (6)–(8).

2.2. Multiple Attributes Heuristic Algorithm for Size-Reduction

Given the complexity of the proposed clustering approach, the goal in this subsection was proposing a heuristic size-reduction algorithm to handle the issue. The present MILP model can be also applied to long timeframe planning, including multiple attributes. However, the linearity and programming basis of the full-scale clustering model was kept by the heuristic modeling framework. As shown, the clustering model was namely composed of two variable types: continuous ($AD_{a,d,h}$, $RV_{a,h,d,c}$, and $D_{a,c,h}$), and discrete (day assignment binary variable $x_{d,c}$). Accordingly, the algorithm decomposed the original problem into a master problem and subproblem. The master problem was an MIP that solves complex variables such as day assignment ($x_{d,c}$) and fixes them to given feasible integers. The subproblem was Linear Programming (LP) that solves the resulting continuous cluster curves problem ($D_{a,c,h}^n$) using the master problem's fixed integer variable values.

In the master problem, the initial guess clusters were fixed while the optimization algorithm solved for the day assignment ($x_{d,c}$). Then, the master problem's solution was used to initialize the subproblem to find a solution for the cluster curves. Hence, turning the problem into a simple linear programming. The algorithm worked on an iterative

structure comparing the upper and lower bound solutions until the differences were within the acceptable range. This structure has been used earlier, and denotes a suitable solution method to deal with large-scale mathematical models [34,35]. It is worth mentioning that the objective function upper and lower bounds were obtained from solving the master problem and subproblem, respectively.

Figure 3a,b shows the multiple attributes heuristic algorithm flow diagram. For instance, Figure 3a illustrates the algorithm execution for a given weight factor combination and each number of clusters. In addition, it shows the execution of the proposed algorithm for a given weight factor to construct the Pareto frontier. After every scenario of a given weight factor has been considered, the procedure moves to the following weight factor repeating the steps until all weight factors are considered.

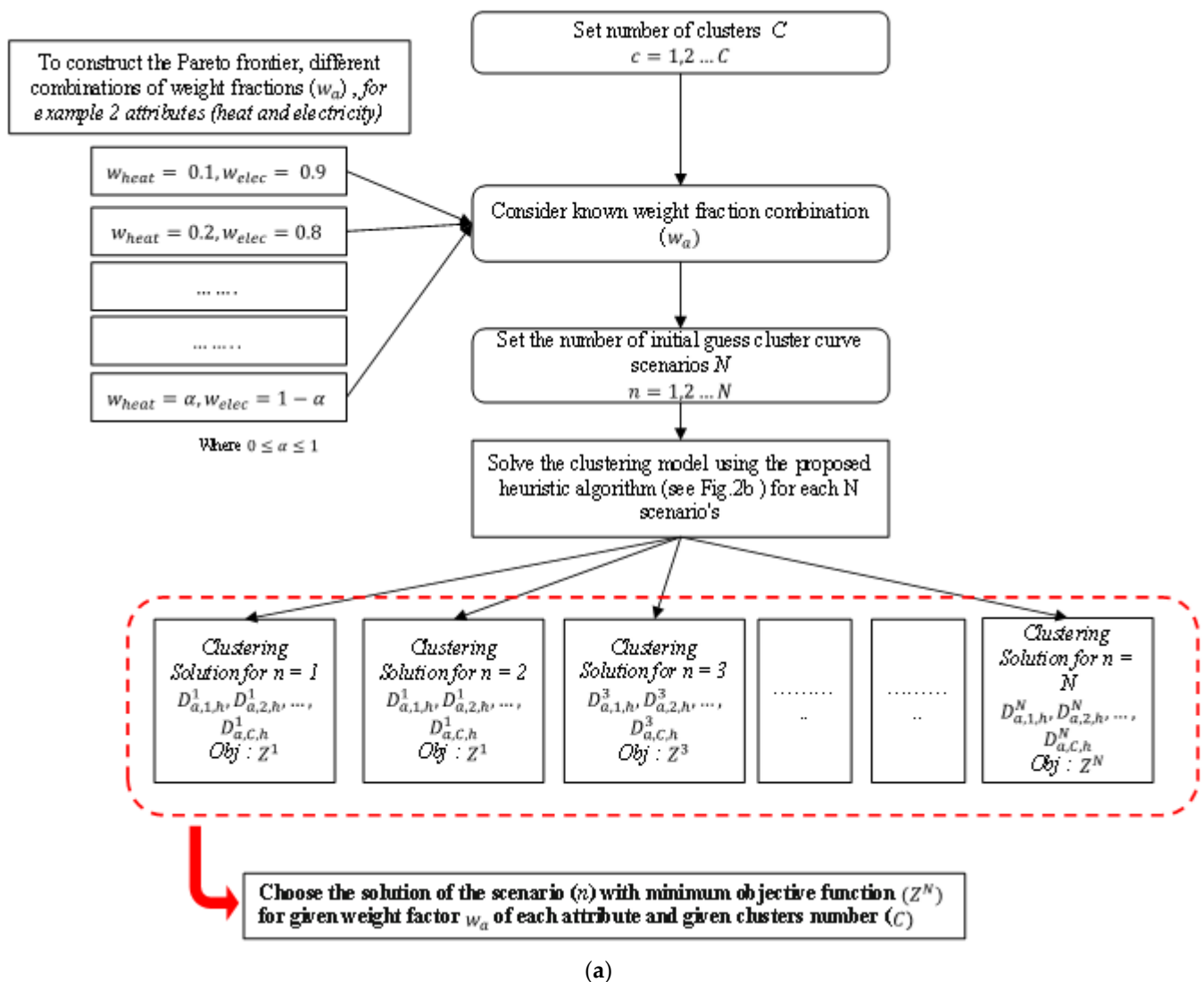
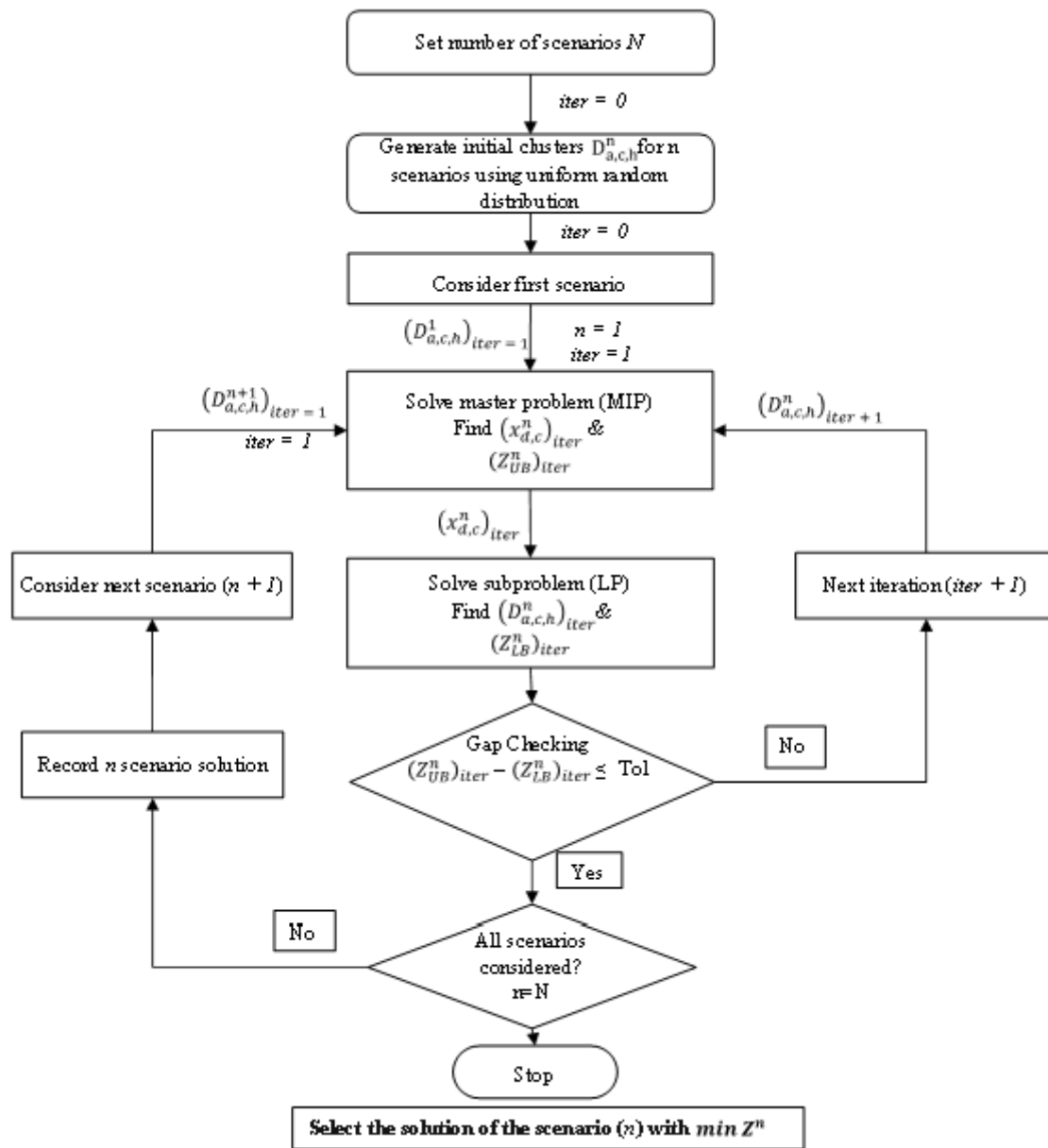


Figure 3. Cont.



(b)

Figure 3. (a) Multiple attributes heuristic algorithm flow diagram for the given weight factor combination and Pareto frontier construction, (b) multiple attributes heuristic algorithm flow diagram for each initial guess scenario.

On the other hand, Figure 3b shows the execution of the proposed algorithm for each initial guess scenario. This procedure is given as follows:

1. **Initialization:** Set the number of initial guess scenarios N .
2. **Generate random initial guess clusters scenarios:** The scenarios are generated using random uniform distribution between the minimum and maximum demand of each hour for every attribute in the entire demand curve $\{D_{a,c,h}^{n=1}, D_{a,c,h}^{n=2}, \dots, D_{a,c,h}^{n=N}\}$.
3. **Initial scenario:** Consider scenario $n = 1$.
4. **Master problem solution:** Solve for the day assignment $(x_{d,c})$ given fixed cluster curves to obtain the upper bound objective function $(Z_{UB}^n)_{iter}$.

5. **Subproblem solution:** Solve for cluster curves $(D_{a,c,h}^n)$ given fixed day assignment $(x_{d,c})$ from the master problem and obtain the objective function lower bound $(Z_{LB}^n)_{iter}$.
6. **Convergency check:** If $|(Z_{UB}^n)_{iter} - (Z_{LB}^n)_{iter}| \leq Tolerance$ go to 7. Otherwise, implement cluster curves $(D_{a,c,h}^n)_{iter+1}$ obtained from subproblem as starting point to solve master problem. Repeat steps 4–6 and iterate until convergence is achieved.
7. **Next scenario:** Record n scenario solution and consider next scenario. Repeat steps 4–6 until all scenarios have been considered. Then move to next step.
8. **Scenario with minimum objective function:** A cluster solution corresponds to the minimum objective function value ($\min Z^n$). The model can be used for both normal clustering (1)–(2) and (4)–(5), or sequence clustering (1)–(2), (4)–(5), and (6)–(8). The common formulation can be applied to multiple attributes problems. The models were formulated in the General Algebraic Modeling System (GAMS) [36]. The total number of continuous and binary variables for the full-scale clustering model with 2 attributes was given as $2 \times 24[D(1 + C) + C]$ and $D \times C$, respectively. The variables D and C denote the total number of days and clusters, respectively. Likewise, the total number of binary variables in the master problem was given as $D \times C$ while the total number of continuous variables in the subproblem was $2 \times 24[D + C]$.

The solution quality and time for the full-scale general clustering and size-reduction heuristic algorithms for multi-attributes were examined. One can notice that the full-scale clustering could tackle the entire year's heat and electricity demand data using normal clustering. For instance, for one year of demand data, no solution was returned after 48 h of CPU time. On the other hand, when using sequence clustering, the solution time was reasonable for the one-year data. In sequence clustering, there were extra sets of constraints that reduced the feasible region size, which resulted in shorter solution times.

For comparison purposes, the two proposed algorithms were tested using reduced datasets (i.e., 20 days) of one year demand data. The runs included 4, 5, and 6 clusters using 20-day demand data for the normal and 365-day demand data for the sequence clustering, respectively (i.e., 6 runs in total). For all runs, the weight factor was set to 0.5 in both attributes. Twenty-five initial guess scenarios were generated in each of the heuristic formulation runs. Table 2 shows the optimal objective function value and solution time for both clustering methods.

Table 2. Computational performance of heuristic and full-scale algorithms.

		Objective Function (MWh)		Solution Time (min)	
		Heuristic Formulation	General Formulation	Heuristic Formulation	General Formulation
Normal clustering—20 days	4	10.836	10.836	1.05	50.25
	5	9.192	9.192	1.23	148.03
	6	8.356	8.340	1.38	434.75
Sequence clustering—365 days	4	469.144	469.144	149.58	469.86
	5	430.696	430.969	500.42	2142.48
	6	404.900	404.54	770.42	11,703.16

As shown in Table 2, the heuristic approach significantly reduced the CPU time compared with the full-scale model solution. As the number of clusters increased, the heuristic algorithm did not match the full-scale model's objective function value. Nonetheless, the difference was negligible. For example, for 6 clusters, the IAE difference between the heuristic and full-scale model was less than 1%. Therefore, the heuristic approach outperformed in terms of the solution time, especially for large datasets with close proximity to the full-scale model's solution form.

Accordingly, we implemented the heuristic approach to generate cluster curves for the energy hub case study presented in Section 3.

2.3. Uncertainty of Wind Speed Modeling

Unfortunately, wind speed is notoriously variable, varying substantially throughout a day, from season to season, and even from year to year. Nonetheless, the Weibull distribution was favorable for describing the wind speed fluctuation at any time interval using two parameters [37]. This statistical tool reflected how often the winds of different speeds could be seen at a location. This was a widely used method in both industry and academia. Therefore, the Weibull distribution was used to fit one-year wind speed data [36]. The wind data corresponded to the measured wind speeds from the Waterloo region in 2018, collected from the National Solar Radiation Data Base [38]. The maximum likelihood method (MLM) was used to fit a Weibull distribution to measured wind speed data [39]. Accordingly, the best-fit Weibull distribution of the available data was achieved. Figure 4 shows the best-fit Weibull distribution and variable wind speed probability. The probability can be estimated as follows:

$$prob(v) = \frac{k}{c} \left(\frac{v}{c}\right)^{k-1} \exp\left[-\left(\frac{v}{c}\right)^k\right] \quad (9)$$

where v is the wind speed, k denotes the shape, and c is the scale.

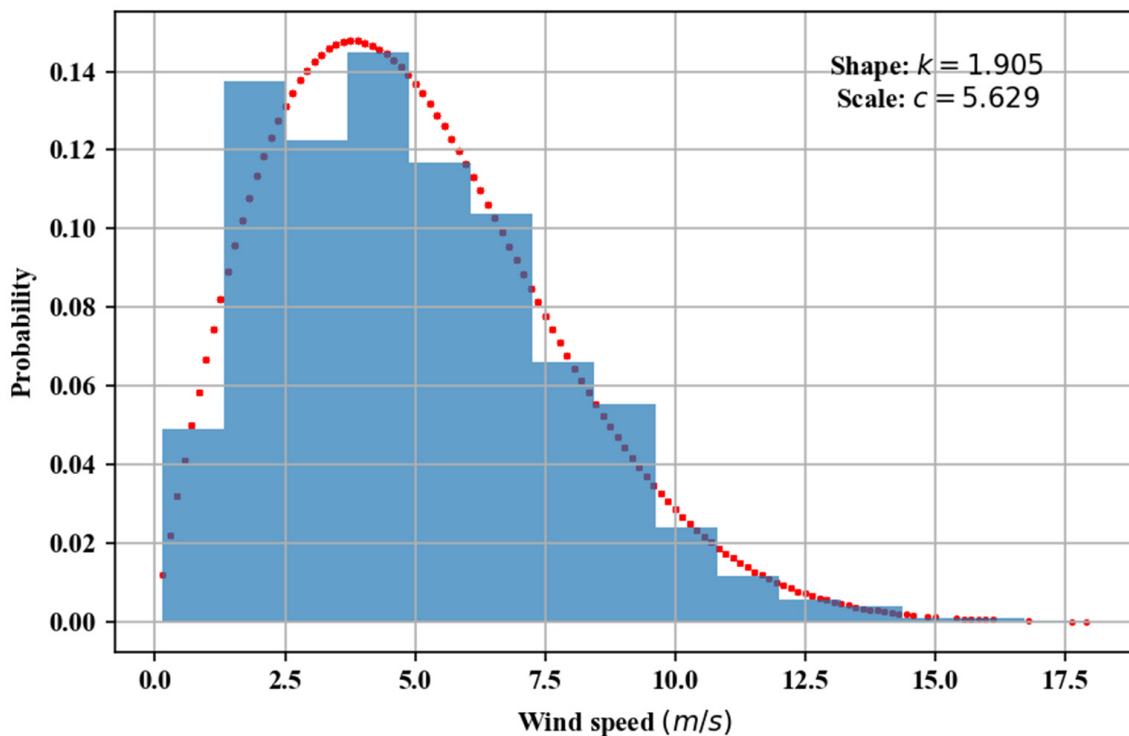


Figure 4. Actual and best-fit of the wind speed profile distribution.

The Weibull distribution allows for estimating the probability of the wind speed occurrence. Accordingly, stochastic scenarios can be generated as each scenario has an occurrence probability (see Figure 5). The probability at which the wind speed is delimited between two points is given as follows:

$$prob_s(v_s^u > v > v_s^l) = \int_{v_s^l}^{v_s^u} \frac{k}{c} \left(\frac{v}{c}\right)^{k-1} \exp\left[-\left(\frac{v}{c}\right)^k\right] dv = \Phi(v_s^u) - \Phi(v_s^l) \quad (10)$$

where Φ is the cumulative distribution function, and v_s^u and v_s^l are the upper and lower wind speed limits of each stochastic scenario (s). Accordingly, the inverse cumulative Weibull distribution (Φ^{-1}) returns the wind speed at a given probability of occurrence.

Additionally, the upper and lower wind speed limit of each stochastic scenario can be calculated as follows:

$$\Phi^{-1}(\text{prob}[v_s \geq v]) = v_s \quad (11)$$

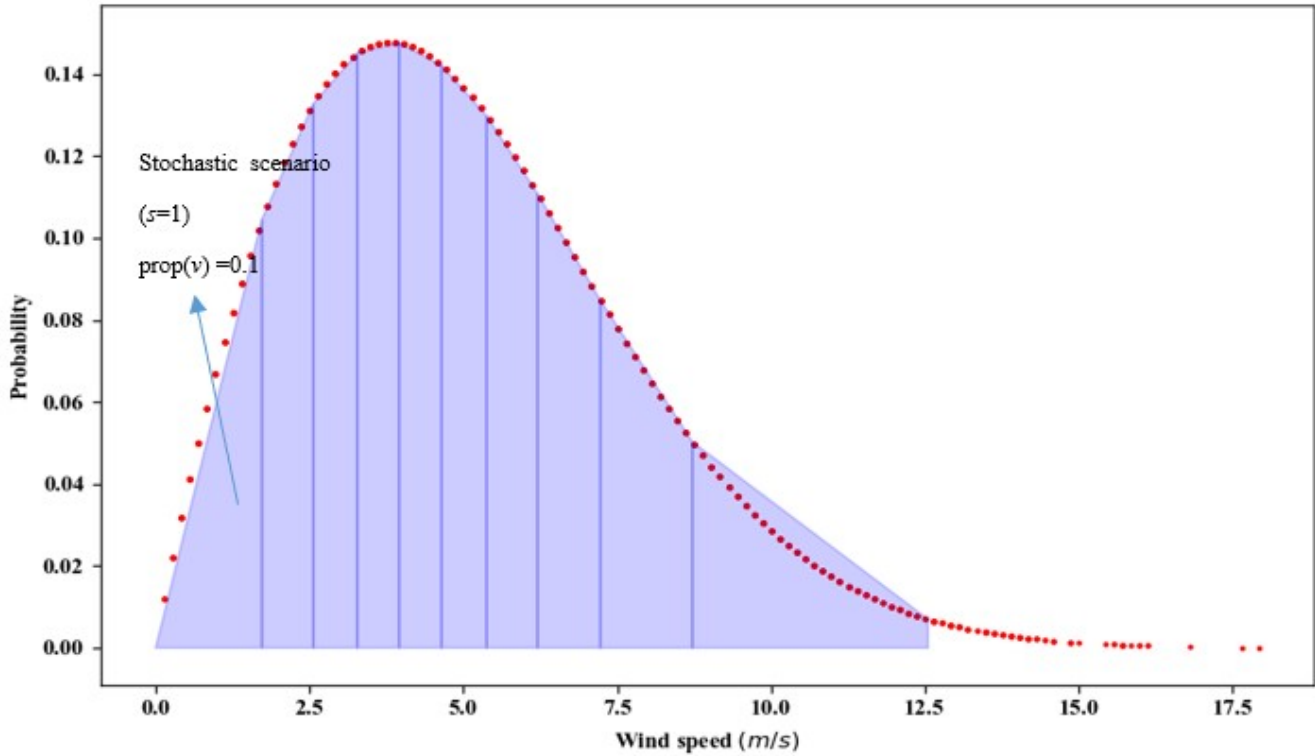


Figure 5. Wind speed stochastic scenarios.

In order to obtain scenarios with equivalent probabilities (i.e., matching areas under the probability density function curve), each scenario is denoted by a probability equal to $\left(\frac{1}{S}\right)$. The variable S represents the total number of scenarios. The upper and lower limits of the wind speeds for each scenario can be calculated as follows:

$$v_s^u = \Phi^{-1}\left(\frac{s}{S}\right), v_s^l = \Phi^{-1}\left(\frac{s-1}{S}\right), \quad \forall s < S \quad (12)$$

$$v_s^u = \Phi^{-1}(0.99), v_s^l = \Phi^{-1}\left(\frac{s-1}{S}\right), \quad \forall s = S \quad (13)$$

To avoid infinite values, Equation (13) is used to impose a 99% confidence interval given that when $s = S$, $\Phi^{-1}(1) = \infty$. Figure 5 illustrates the fitted Weibull distribution curve where each shaded area under the probability distribution function represents a single stochastic scenario with the probability of $(1/S)$.

3. Case Study: Multiple Attributes Clustering Application to Energy Hubs

The present case study illustrates the application of the clustering algorithms to energy demand data including multiple attributes, and data-driven statistical methods to represent the intermittent behavior of uncertain wind speed data. Additionally, the energy hub design and operation were formulated as a multiscale model with multiple attributes. This was done by agglomerating demand data with similar profiles and generating stochastic scenarios for a two-stage stochastic model considering uncertainty in the wind data. In addition, the impact of clustering on the solution accuracy was investigated. It has previously been determined that clustering considerably helps to reduce the computation time.

The present case study evaluated the proposed sequence and normal clustering algorithms' outputs versus the full-size energy hub design and operating model under wind speed uncertainty with multiple demand attributes. The energy hub is a strategic (long-term) and medium-term decision level problem. The aim was to minimize the total annual cost of designing (installing and sizing) and operating the energy hub while meeting the energy demands. There are numerous models available in the literature for the energy hub problem, from mathematical programming to heuristics. The present case study adopts the [40] MILP model.

The present energy hub system aimed to minimize the annual operational and maintenance cost, as well as the capital cost while meeting electricity and heat demands within the units' operating capacities and physical constraints.

In this paper, the authors consider both REDSs (Renewable Distributed Energy Sources) and DERs (distributed energy resources) based on fossil fuels. The current energy hub system includes a variety of conventional energy conversion technologies powered by natural gas such as combined heat and power (CHP) units, boilers, and a non-conventional energy conversion technology (i.e., wind turbines) powered by renewable energy resources. Additionally, it utilizes a hydrogen production and storage system from electricity utilizing an electrolyzer, hydrogen tank, and fuel cell as the ESS. We chose a hydrogen storage system because it can play a role in both storing energy and supplying the hydrogen demand for hydrogen vehicles. Figure 6 shows the energy hub layout with the considered energy technologies and input data handling (wind speed, electricity, and heat demand).

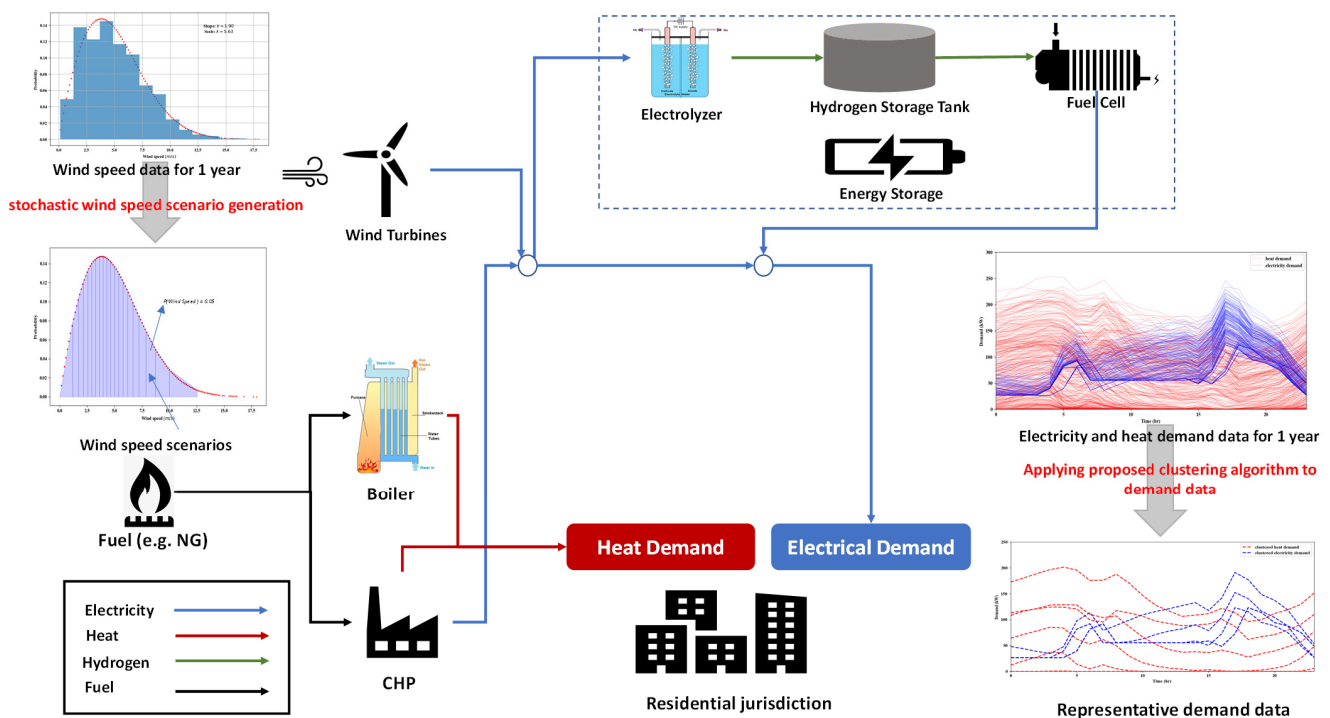


Figure 6. Understudy energy hub architecture.

The optimization program decides the number of each unit and the respective capacity within the energy hub system, as well as the operating points for the electrolyzers, hydrogen tanks, fuel cells, boilers, and CHP generators at each time point. Particularly, in this paper, the discrete size of each technology was considered in the optimization, which made this work more realistic. The number and type of each technology chosen was a design decision variable while the operating variables were related to how the energy hub units were running. The main outputs of the optimization model can be summarized as follows:

1. Type and number of energy conversion and storage technologies within the hub.

2. Design and operation of optimal energy hub under intermittent wind energy availability, and based on multiple attributes aggregated demand data or full-size demand data.
3. Economic cost of the system including capital, operation and maintenance, and fuel consumption.
4. Environmental impact of the system through the GHG emissions.

Natural gas is the fuel for both: the boiler and CHP. As illustrated, the electricity demand was met by means of the CHP generators, wind turbines, and fuel cell, whereas heat is met by the boilers and CHPs. The list of the energy generation technologies and its technical and economical properties are given in Table 3. This model is a general framework for microgrid/energy hub systems where different technologies can easily be added or removed, according to the problem that needs to be solved.

The mathematical model was formulated as a two-stage stochastic with recourse, where the first-stage decisions decided the design of the system that included the number of each unit and the respective capacity within the microgrid, while the second-stage decisions planned the operation of the system including the operating points for the electrolyzer, fuel cell, CHP units, and boilers at each time point. The two-stage stochastic recourse (we refer to it as recourse problem, RP) formulation was basically a bi-level optimization formulation whose inner optimization problem mimicked the second-stage planning process. Due to the special structure, the two-stage stochastic programs could be naturally reformulated into an equivalent single-level optimization problem. Therefore, the single level optimization formulation of the RP for the design and operation of the energy hub system could be directly written as follows:

Table 3. Technical and economic information about the energy conversion and storing technologies.

Unit	Rated Capacity (kW)	Input Energy Form	Output Energy Form	Efficiency	Capital Cost	Operating and Maintenance Cost (\$/kW)
Boiler [9]	530	kW fuel HHV	kW heat	0.82	100 (\$/kW)	0.027
	300			0.9	120 (\$/kW)	0.027
	100			0.8	150 (\$/kW)	0.027
CHP [9]	300	kW fuel HHV	kW power	0.26	900 (\$/kW)	0.016
	100	kW fuel HHV	kW heat	0.44		
			kW power	0.35		
60	kW fuel HHV	kW heat	0.5	1080 (\$/kW)	0.016	
wind turbine [41]	20	kW available by air	kW power	0.4	2200 (\$/kW)	0.008
	30	kW available by air	kW power	0.42	1906 (\$/kW)	0.008
Electrolyzer [9]	Storing units		kg of H ₂	0.0193	155,051\$/unit	0.06
	290	kW power				
Fuel Cell [42]	250	kg/h of H ₂	kW power	16.5	210,630\$/unit	0.06

The detailed equations of the stochastic energy hub model are presented in Section S2 of the Supplementary Information.

Key case study inputs such as the heat and electricity demand in addition to the number of cluster curves have been previously generated and can be found in Section S1 of the Supplementary Information. Moreover, the objective cost function was multiplied by a parameter designated as γ_d (see Equation (14) for details) which allowed comparing of the original demand dataset (i.e., one-year time horizon) and the clustered cases. The parameter γ_d denotes the repetitions (frequency) for the d^{th} day. The parameter γ_d is equal to one when the original demand data is used, and equal to the number of days when the representative cluster curves are used. For instance, if cluster 1 represents 45 days, its corresponding parameter γ_d is equal to 45.

$$\min_{CRF} \left[\sum_u y_u CAP_u E_{max}^u + \sum_{wt} y_{wt} CAP_{wt} \sum_{st} y_{st} CAP_{st} \right] + \sum_s \beta_s \left[\sum_d \gamma_d \left(\sum_h \left(\sum_u \left[P_{i,d,h,s}^u OM_u + NG_{d,h,s}^u Price_{ng} \right] + \sum_{st} P_{i,d,h,s}^{st} OM_{st} \right) \right) \right] \quad (14)$$

For future reference, the energy hub with hydrogen storage considering hourly electricity and heat demand loads for an entire year (i.e., full-size) was designated as the original model. On the other hand, the energy hub with hydrogen storage considering 4, 5, and 6 hourly load clusters (i.e., clusters regarded as days) was designated as the clustered model (see Section S1 of the Supplementary Information for more details). Moreover, Table 4 lists the weight factor combinations employed to construct the Pareto frontier.

Table 4. Weight factors for the multi-objective function.

Weight Factor	Heat	Electricity
1	0.8	0.2
2	0.7	0.3
3	0.6	0.4
4	0.5	0.5
5	0.4	0.6
6	0.3	0.7
7	0.2	0.8
8	0.1	0.9

This case study comprised four different scenarios. Each scenario considered a particular operating or environmental constraint under which the proposed clustering algorithms and data-driven methods were tested and evaluated. Accordingly, the following four subsections present the results and discussions of each of these scenarios.

3.1. Baseline Scenario

This case study considered the energy hub operation under unconstrained GHG emissions. For instance, Figure 7 shows a comparison of the objective function values along with their corresponding relative errors for the clustered and original models (optimal). As shown in the figure, all clustered cases underestimated the objective function value compared with the original case. The objective function values were closer to the original model in normal clustering compared with sequence clustering. The clustered model's objective functional relative error (i.e., compared with the original model) ranged between -4% and -10% . Moreover, the higher the clusters number, the better the quality of the solution for both sequence and normal clustering. Thus, the solution gap between the clustered and original case became smaller.

Regarding the relative errors, the average relative error of all weight factors (see Table 4) is included in Figure 7. The errors were inversely proportional to the number of clusters. The clustered model's objective function values did not significantly vary as a function of the weight factors due to symmetry similarity in the electricity and heat demands. The bar chart (y-logarithmic scale) in Figure 7 displays the average solution time for all weight factors of each clustered case run (4, 5, and 6 clusters) versus the original model solution time. As shown in the figure, the clustered model significantly outperformed the original model in terms of the solution time. The clustered model's average solution time was two orders of magnitudes smaller than the original model (i.e., ~ 7000 s).

In addition, we examined the effect of applying the multiscale clustering approach for the demand data on the energy hub design. Figure 8 showcases the design decision variables of each energy hub unit and the total installed heat generation and electrical power capacity. This is presented for the clustered and original model for the weight factors 1, 4, and 8, using sequence and normal clustering. The figure illustrates that the higher the clusters number was, the closer the design decision variables' values between the clustered and original model were. Likewise, the installed generation capacities followed similar trends. It is worth noticing that overall, the weight factor 1 showed slightly better results for normal clustering because it tended to better align with the heat demand. Due to the high fluctuations in the heat demand, an improved design (i.e., closer to the original model)

was attainable by prioritizing the heat demand, which minimized the errors caused by cluster variability.

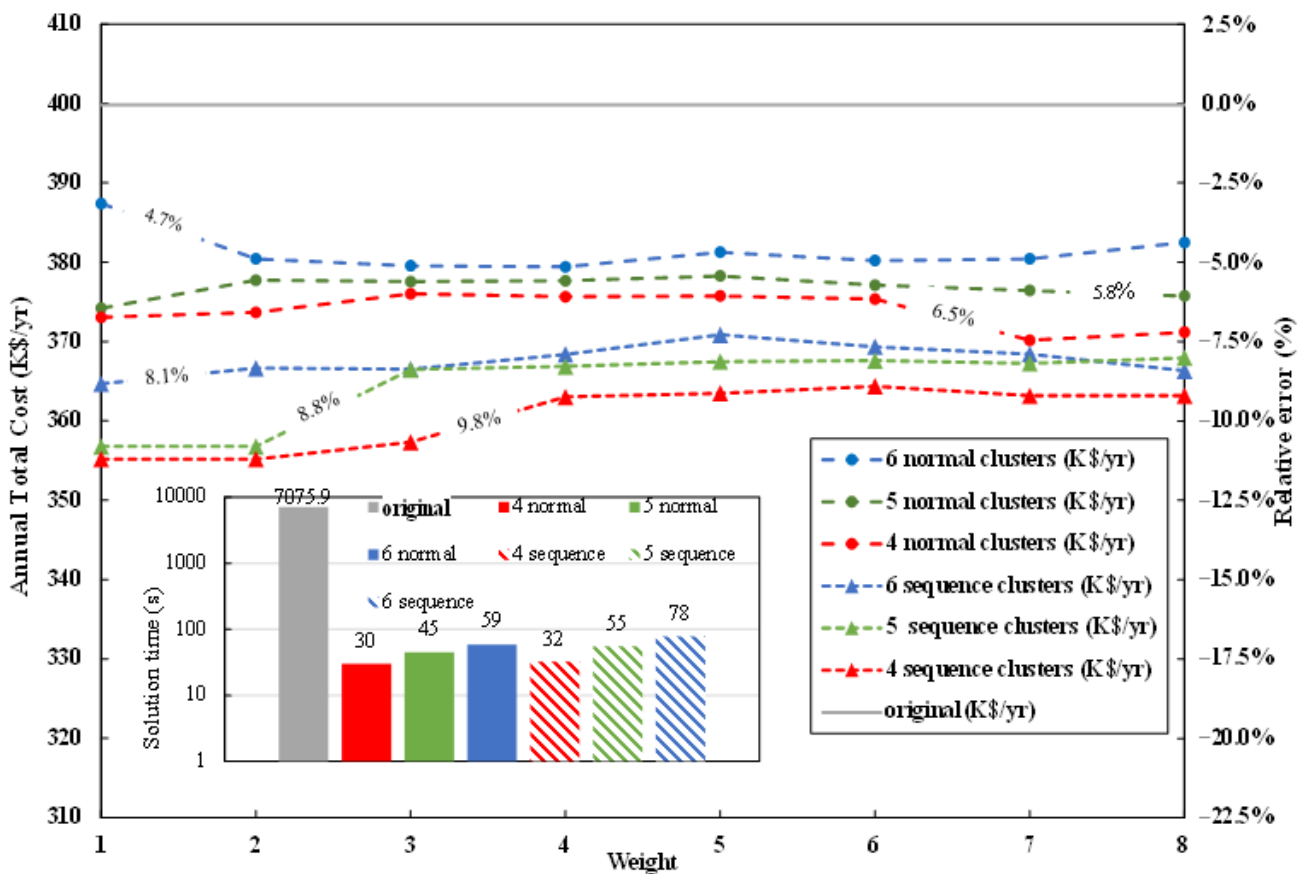


Figure 7. Comparison between original and clustered energy hub solutions in terms of solution quality and time.

As shown in Figure 8, the clustered cases’ installed capacity for power and heat generation are generally underestimated. Specifically, the power capacity was underestimated by a lower margin than heat compared with the original model. This was because the total heat production rate was allowed to exceed the demand (if necessary), whereas an equality constraint was imposed on the power balance to satisfy the electricity demand. Generally, changing the weight factors had a steady effect on the installed capacity of power and heat, as the priority switched from heat to electricity. This could be the result of the heat and electricity demand featuring similar symmetry throughout the whole horizon.

3.2. Environmental Scenario (CO₂ Emissions Regulation)

The previous results showed that the optimization emphasized non-renewable energy sources, and not a single wind turbine storing unit (e.g., electrolyzer and/or fuel cell) was installed. This was the result of traditionally higher costs of renewables compared with traditional fossil fuels. Nonetheless, renewables are cleaner alternatives which can be integrated with current energy hubs and microgrid systems to mitigate GHG emissions (i.e., CO₂, CH₄, NO_x). To analyze the energy hub design under CO₂ emission regulations, a carbon constraint was introduced in the energy hub mathematical model as follows:

$$Em = \sum_s \beta_s \left[\sum_d \gamma_d \left(\sum_{u,h} (\delta * b * NG_{d,h,s}^u) \right) \right] \leq \alpha \tag{15}$$

where E_m is the annual CO₂ emission mass generated by the energy hub, δ denotes Ontario’s natural gas emission factor (0.187 kg/kWh), and α is the CO₂ emissions limit. Only emissions from the operation of fossil fuel units (boilers and CHP) were considered, whereas emissions associated with renewable (i.e., wind turbines) and storage units were considered negligible. Renewable units’ emissions in the operation stage were negligible compared to fossil fuel units.

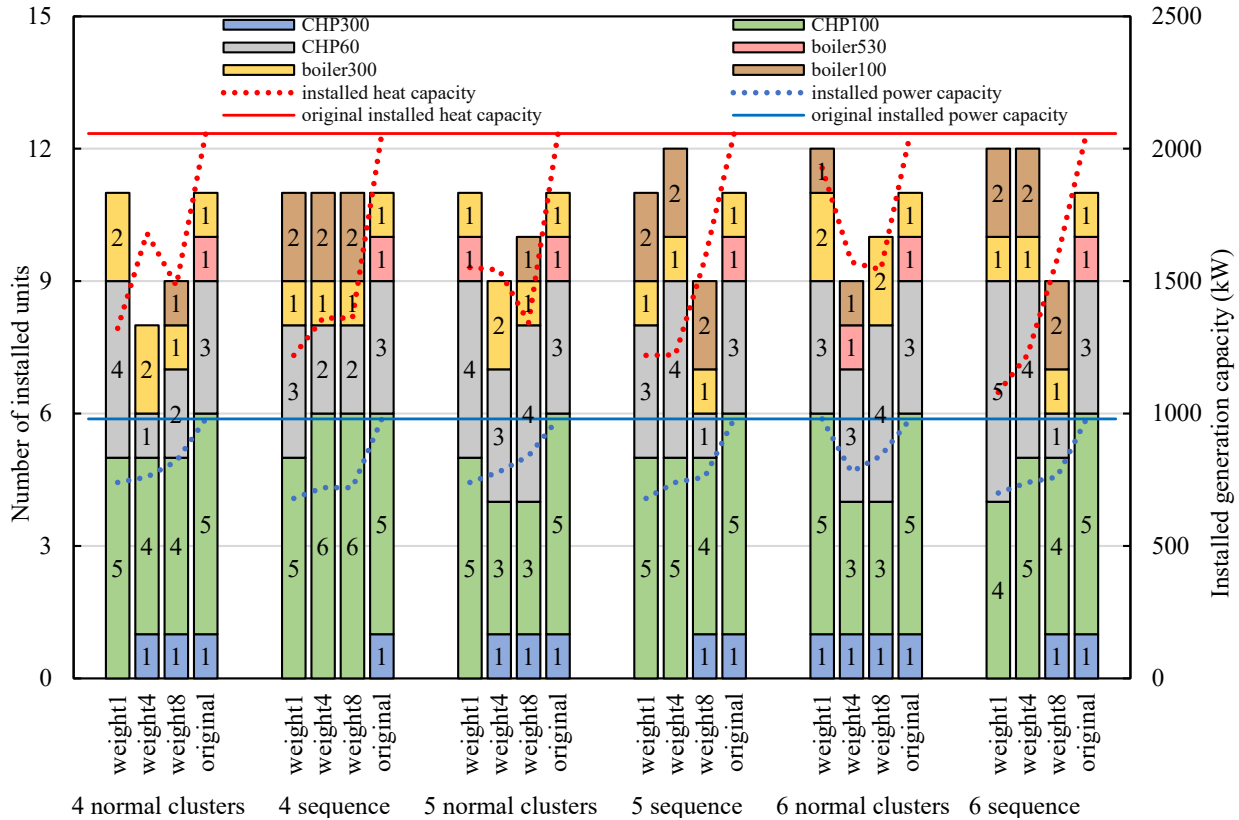


Figure 8. Design results comparison between original and clustered energy hub models.

A sensitivity analysis was performed to assess the effect of introducing carbon emissions regulations. The analysis consisted of fixing the allowable annual CO₂ emissions of the energy hub. Accordingly, Figure 9 shows the energy hub’s total annual cost and installed wind turbine units as a function of the CO₂ emissions. The figure was generated using the clustered model for 6 clusters and a weight factor of 4 (i.e., equal emphasis of heat and electricity data) using normal and sequence clustering. Both conditions better represented the entire year demand data as they featured lower IAE. The upper CO₂ emission limit coincided with the lowest annual cost. At this point, the emission constraint was inactive and the installed wind turbines were nil.

On the other hand, when α was reduced, the objective function (total annual cost) increased and the optimization selected wind turbines. Accordingly, the greater the CO₂ emissions reduction, the higher the number of installed wind turbines and objective function value. It is worth noticing that the results trend for both normal and sequence clustering were nearly equivalent as a function of α . Nevertheless, at higher CO₂ emission levels, the sequence clustering tended to feature slightly lower objective function values and fewer wind turbines. Conversely, at lower CO₂ emission levels sequence clustering chose to install higher numbers of wind turbines leading to total costs exceeding the normal model solution.

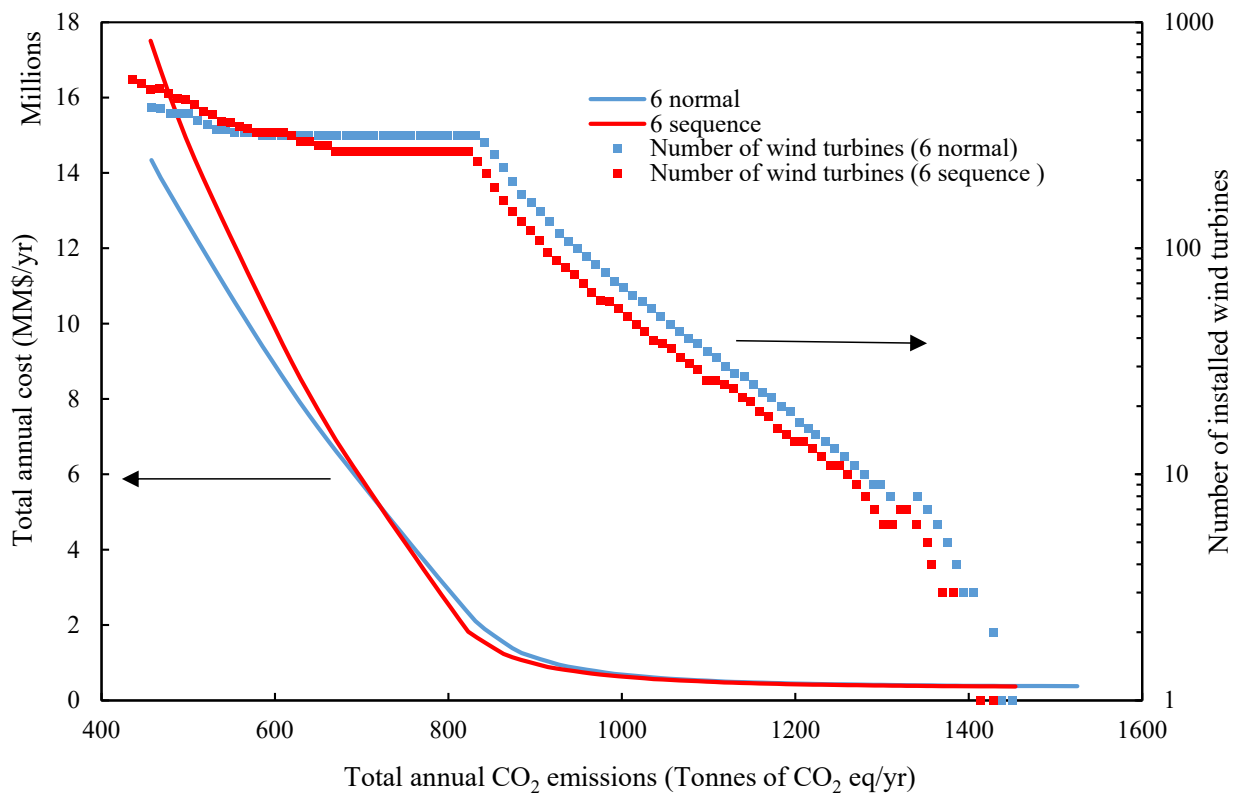


Figure 9. Effect of CO₂ emissions regulation on the objective function (lines) and number of wind turbines (square marker).

Likewise, at lower CO₂ emission levels, the greater the number of recommended storing units for sequence clustering were compared with normal clustering. Emission reductions at high carbon levels slightly impacted the objective function value. Conversely, further emission reductions at already low CO₂ emission levels came with moderate cost increases. These additional costs arose from the extra storage units required to help dispatch wind power more efficiently.

3.3. GHG Emissions Constrained Scenario

This scenario assumed a 20% CO₂ emissions reduction from the upper carbon limit (i.e., baseline scenario when the emissions constraint is inactive). The effects of the weight factors and clusters numbers over the objective function value as well as the relative errors are illustrated in Figure 10. For instance, when the clusters were emphasized more on the heat demand (at weight factor 1), the clustered model's objective function values were much closer to the original model. This was because the heat demand undergoes a higher degree of variability among utilities. When clusters prioritize electricity demand in normal clustering (at weight factor 8), the highest deviation or relative error takes place with respect to the original model results. The proposed weighted clustering method allows tuning and generating clusters that prioritize attributes over others.

There was no clear relationship between the weight factors 1 to 8 and solution quality. Nonetheless, sequence clustering exhibited less variability as the priority switched from heat to electricity. In addition, the average relative errors tended to converge. Normal clustering showed slightly lower average errors than sequence clustering. The average solution time for all weight factors of each cluster run (4, 5, and 6 clusters) along with the original energy hub solution time are displayed in Figure 10. From the bar chart (y-logarithmic scale), it is observed that the time required to solve the clustered cases was notably lower than that of the original model. For instance, the original model's solution time (65,137 s) exceeded the clustered model's average solution time by between an order

of magnitude of 2 and 3 (i.e., 50 to 100 s). It is worth noticing that solving the problem without CO₂ emissions regulation (see Figure 7 for details) was significantly faster by an order of magnitude of 1.

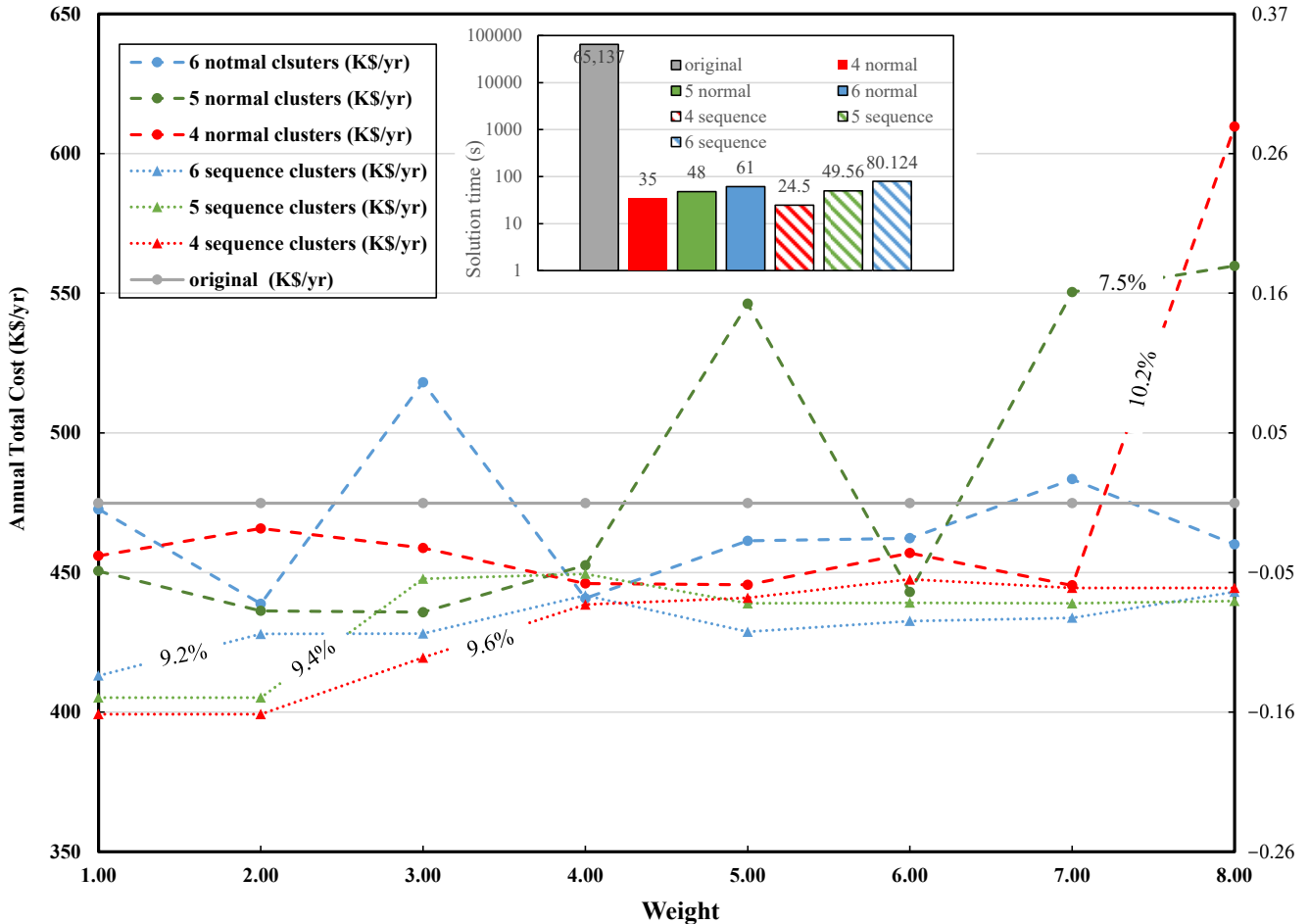


Figure 10. Comparison between the original and clustered energy hub solutions in terms of solution quality and time under the GHG emissions constraint.

On the other hand, once the GHG emissions constraint was active, the optimization algorithm chose storing and wind turbine units to keep emissions within the desirable target. Accordingly, additional non-zero variables (e.g., continuous variables associated with power flow to/or from storing units, hydrogen flow rates, power directed from wind turbines, and binary on/off variables for charging and discharging storing) were handled by the optimization algorithm. This significantly increased the degree of complexity. Comparatively, there was no outstanding difference between the clustered cases with/without an environmental constraint in terms of the solution time.

The effects of cluster numbers and weight factors on the design decision variables under a GHG emissions constraint are displayed through Figures 11–13. For example, Figure 11 shows a comparison between the original and clustered model fossil fuel units’ design variable values as a function of optimization runs for weight factors 1, 4, and 8. As illustrated in the figure, the weight factors had no significant effect on the design decision variable values. Like the previous scenarios, the higher the number of clusters, the closer the solutions are to the original model (see Section S3.2 of the Supplementary Information for more details). Furthermore, all optimized clustered scenarios avoided the selection of CHP300 and boiler530 units. This aligned with the original model’s results, as these types of units are the largest in size and are the greatest carbon emitters.

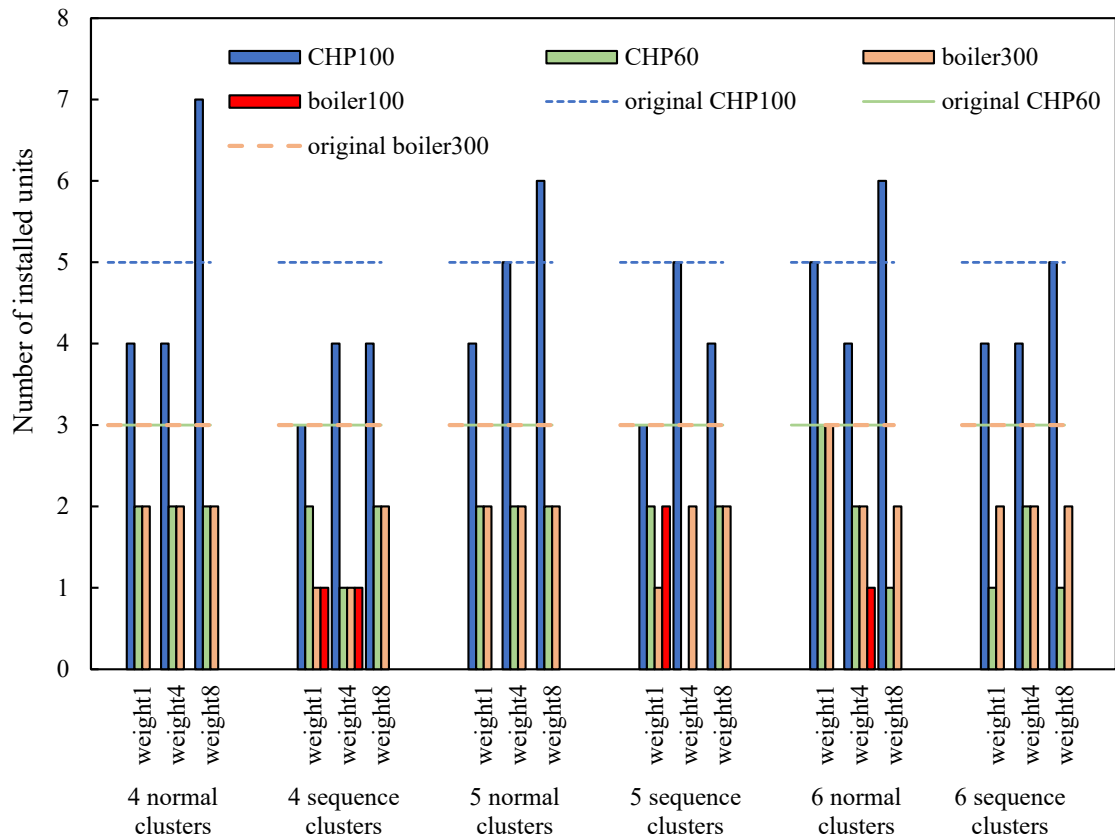


Figure 11. Number of fossil-powered units in the energy hub under the GHG emissions constraint.

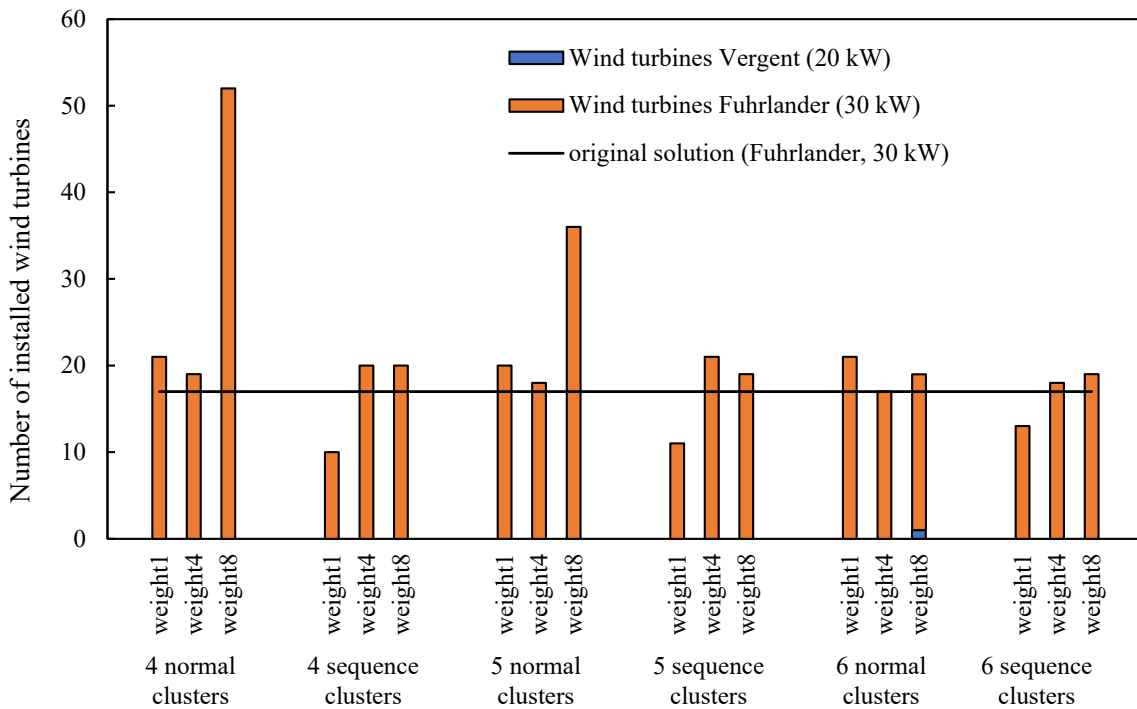


Figure 12. Number of wind turbines suggested by the original and clustered models under the GHG emissions constraint.

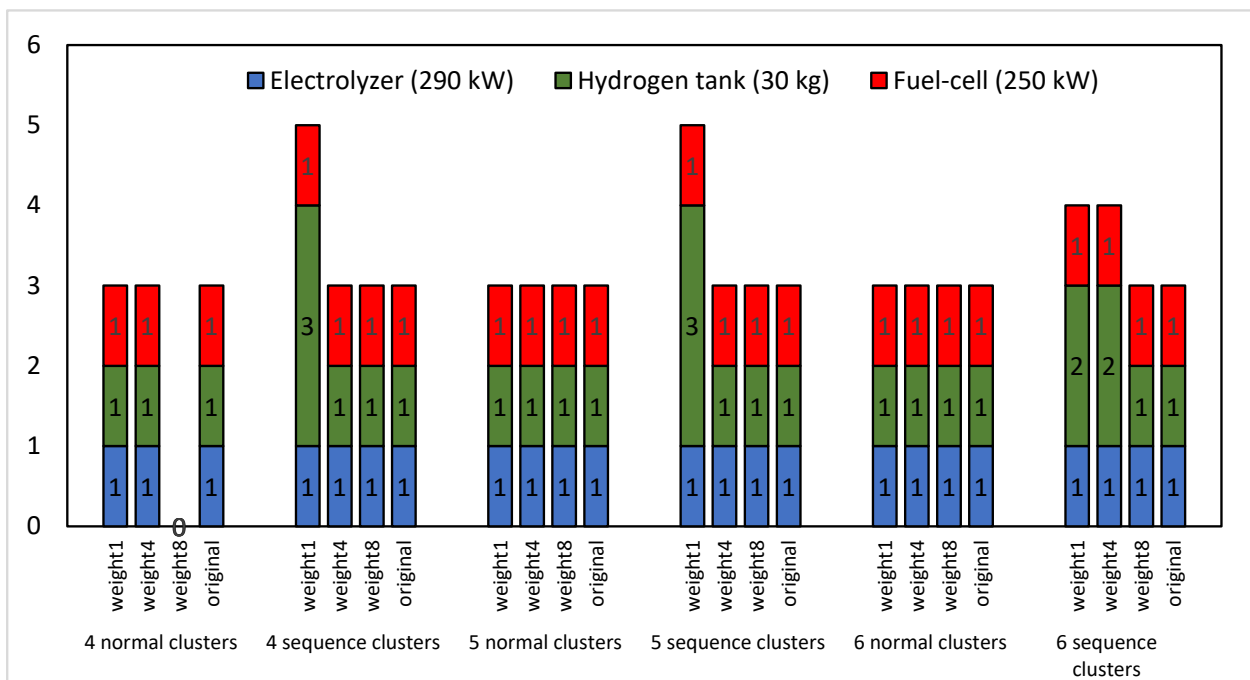


Figure 13. Number of installed storing facilities suggested by the original and clustered models under the GHG emissions constraint.

Figure 12 shows the number of wind turbines versus the clustered runs for the weight factors 1, 4, and 8 (also the original model results). As illustrated in the figure, the larger the number of clusters was, the smaller the gap was in the number of wind turbines between the clustered and original model results. It is also worth noticing that at the weight factor 8 for 4- and 5-normal clustering, the number of wind turbines was overestimated by a larger margin. This was explained by the high error in the objective function value as illustrated in Figure 10.

Figure 13 shows the storing units' design decision variables' values under the GHG emissions constraint using both the clustered and original data. The bar chart shows that for most clustered scenarios, the storing units were in very good agreement with the original model. One can notice that some of the sequence clustering results were overestimating the number of hydrogen tanks needed.

This scenario also discussed the operational/decision solution quality of the proposed multiscale clustering approach of the demand data. Figures 14 and 15 illustrate the energy hub's utility production rates (clustered with weight factors 1, 4, and 8 and original model) by fossil-powered and wind/storage units, respectively. Each unit's utility production was estimated by adding its corresponding production rate over the year. This was the summation of all products between the stochastic scenarios and their corresponding weighted probability. Figure 14 shows that the clustered model's utility production rates are in very good agreement with those of the full-size original model. There was no significant variation in the CHP units' heat and electricity rates between the two models. Nevertheless, the boilers' heat rate relative error was high in sequence clustering.

Figure 15 clearly shows there is a large degree of error in the electricity rates from wind turbines and fuel cells when comparing the clustered and original model results. This deviation from the original model was accentuated in sequence clustering. Despite the errors, the proposed clustering approach could still be considered as a powerful size-reduction tool that was able to reduce the computational time considerably. For example, the design decision variable values between both models were close. Similarly, the heat and electricity production rates (see Figure S13 in the Supplementary Material) from the

clustered model were in very good agreement with the original model results, whereas their corresponding relative errors did not exceed 20%.

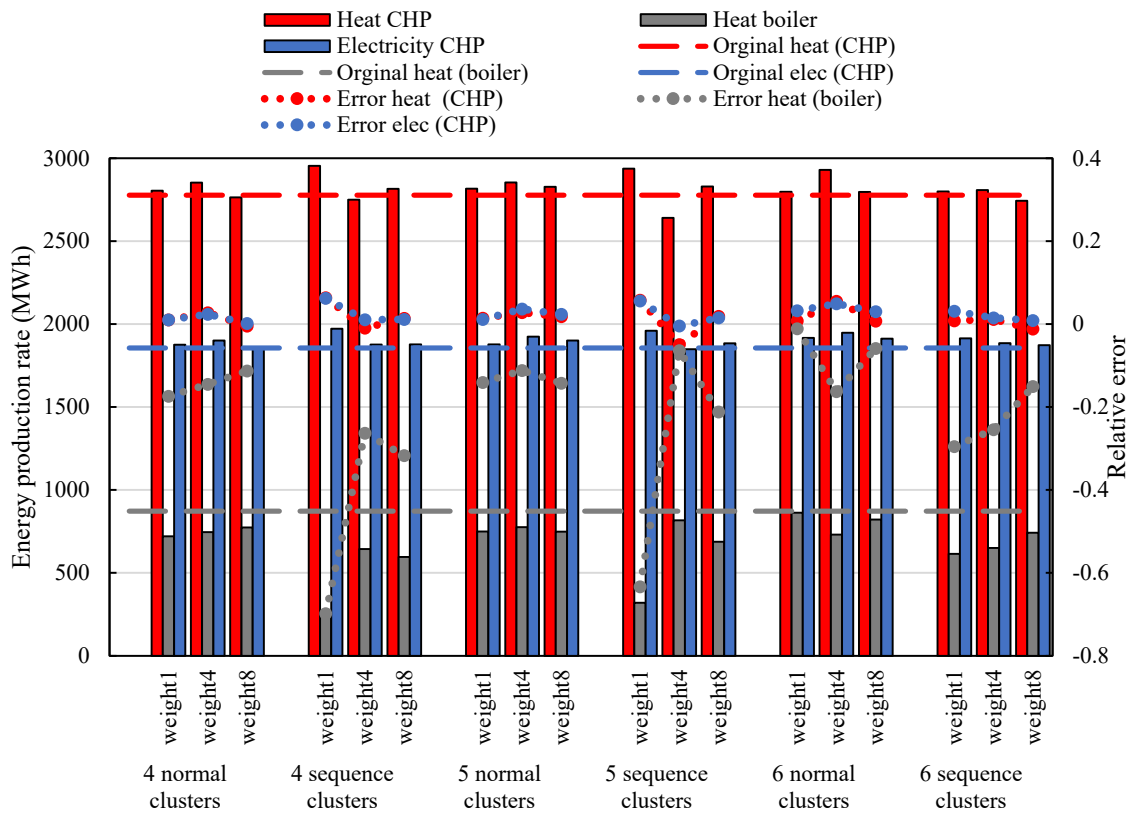


Figure 14. Comparison between the original and clustered models' utility production rates from fossil-powered units under the GHG emissions constraint.

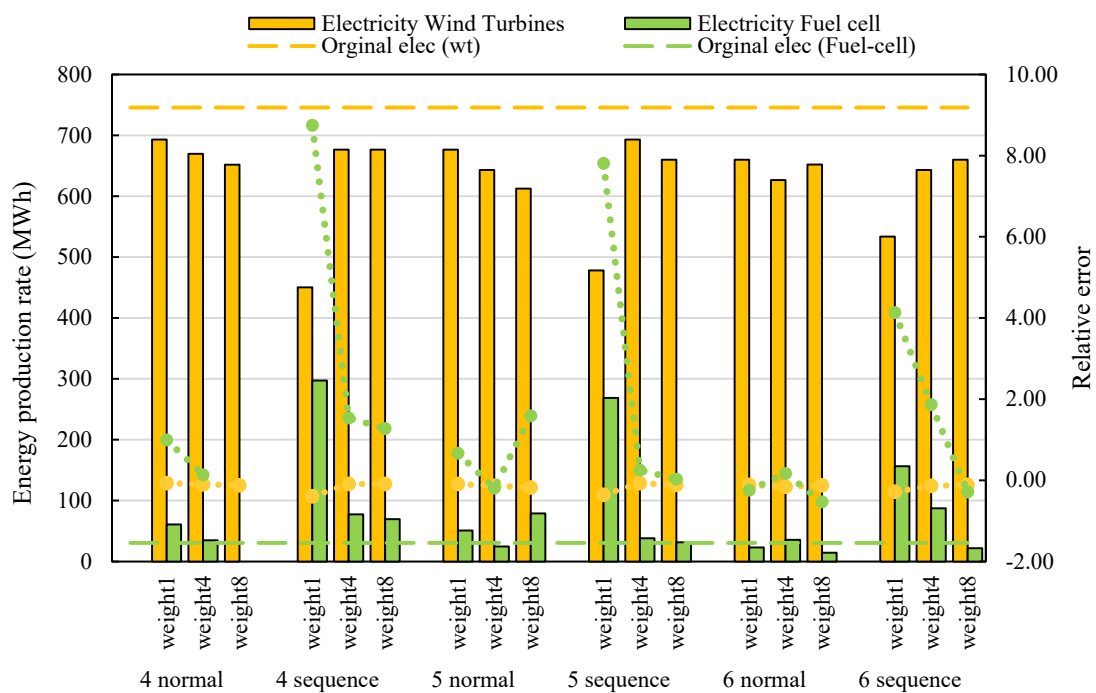


Figure 15. Total utility production by wind turbines and storing units for the clustered and original model under the GHG emissions constraint.

3.4. Stochastic Energy Hub Formulation Assessment

The present section assesses the potential benefits of the proposed energy hub model formulation to store energy under uncertain wind speed scenarios. Accordingly, Figures 16 and 17 illustrate the average power transferred to electrolyzers (i.e., charging power) and received from fuel cells in each stochastic scenario, respectively. Both figures illustrate the results of the energy hub model with 6 normal clusters and a weight factor of 4. For simplicity purposes, the average hourly power over a year per scenario is displayed (average power flow from each hour with respect to all days over a year of time per scenario). As one could expect, Figure 17 demonstrates that the rate of charging (i.e., power transferred to electrolyzers to produce hydrogen) is larger at higher wind speeds. This means that more energy can be stored at a higher wind energy availability. An increase in the scenario number denoted an increase in the wind speed. Additionally, at relatively low demand times, the optimization algorithm decided to store more energy. Conversely, Figure 17 clearly shows that the fuel cells' discharge rate is inversely proportional to the wind speed scenario number. Most discharged power occurred at peak demand.

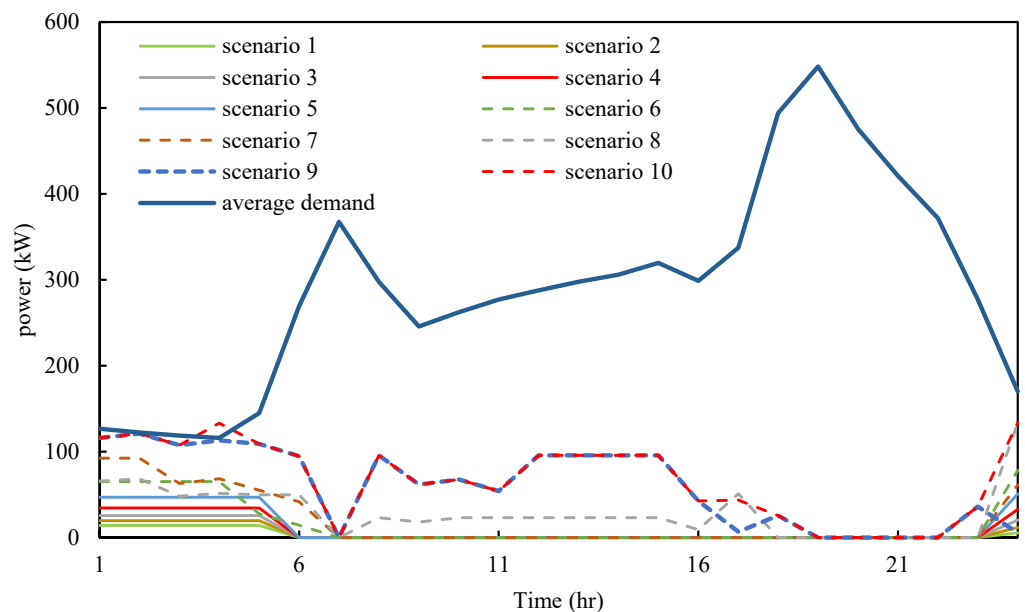


Figure 16. Average charging power per stochastic scenario.

To examine the stochastic programming method efficiency, the value of stochastic solution (VSS) was calculated following [43]. The VSS helped in determining whether it was advantageous fixing the stochastic model's first-stage decision variables. This was done based on the expected value problem (EV) solution. In other words, VSS represented the extra cost the decision maker must pay for not considering uncertainty in the analysis (stochasticity). To estimate the VSS, the solution to the EV must be determined first. In the present study, the EV was represented by the deterministic energy hub solved that considered the mean as the uncertain parameter (wind speed) value. In the next step, the first-stage decision variables (design decision) obtained from the EV were fixed and used as input parameters in the two-stage stochastic energy hub with the recourse problem (RP). Then, the resultant RP was solved after fixing the first-stage decision variables, which was the expected result of using the EV solution (EEV). The EEV provided the second-stage decision variables solution once the first-stage decision variables were fixed. Accordingly, the VSS could be defined as the difference between the EEV and RP.

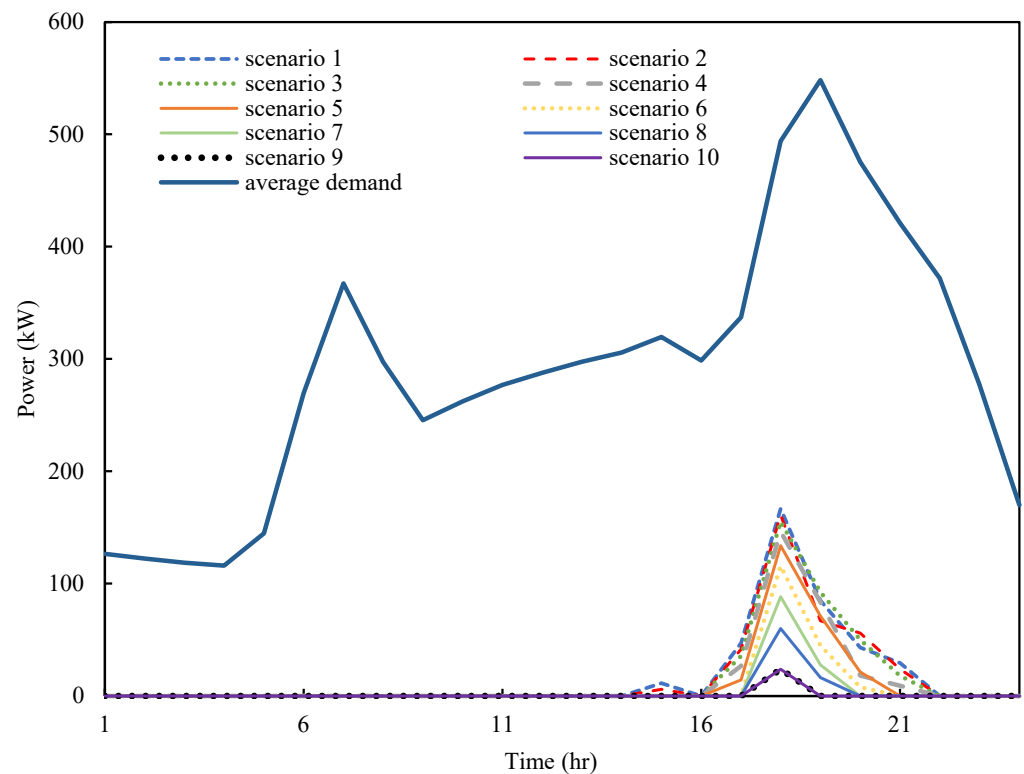


Figure 17. Average discharging power per stochastic scenario.

Table 5 shows the EV, EEV, and RP solutions to the energy hub problem with/without the GHG emissions constraint. The results were obtained using 6-normal clusters with a weight factor of 4 for the year demand data (i.e., better representative by featuring lower IAE). The table clearly shows that when there was no environmental consideration, no benefit was gained from using stochastic programming (i.e., $VSS = 0$). As previously discussed, when the environmental constraint was inactive, neither the wind turbines nor storing units were suggested to be installed; hence, all stochastic scenarios' solutions were identical.

Table 5. Objective function values for EV, EEV problem, and recourse problem (RP).

Objective Function Value:	Results	EV	EEV	RP	VSS
Annual cost (\$/year)	without environmental consideration	379,411.5	379,411	379,411	0
	with environmental consideration	438,717.6	455,561.6	440,729	14,832.7

In contrast, when the emission constraint was active, the VSS was estimated to be 14,832\$/year ($VSS = EEV - RP$). The positive VSS value proved that considering uncertainty in the energy hub modeling is beneficial. Although the EV (deterministic solution) featured the lowest objective function value, deterministic modeling solutions are insufficient because they heavily rely on a relatively small segment of information (e.g., average wind speed). This information does not sufficiently explain real events, such as the wind speed behavior; therefore, they cannot be considered true representatives of real data (e.g., annual wind data). As a result, it could be stated that the wind speed uncertainty has a significant effect on the optimization solution once environmental constraints are considered in the model's formulation (as previously proven by the VSS value estimation).

4. Conclusions

The present work targeted the integrated supply chain problem using a clustering approach. Given the fact that employing shorter time periods (e.g., hours) to achieve optimal decisions leads to larger and intractable integrated models, this work aimed to decrease the model size by representing the yearly days by typical days representative of the operating year. Accordingly, a mathematical programming approach was considered to model the clustering problem with multiple attributes. Distinct attributes featured varying scales or units, which turned the problem into a multi-objective optimization program. Accordingly, the weighting method allowed for dealing with such problem. The present clustering algorithm featured a unique characteristic that enabled attaining normal and sequence clustering employing the same similarity measure. The proposed weighted clustering method allowed tuning and generating clusters that prioritized some attributes among others.

Although the developed approach is simple, the computation complexity of the proposed clustering algorithm is obvious. A heuristic clustering approach was proposed to tackle the computational tractability of the full-scale clustering model. It was found that the larger the number of clusters employed was, the closer the solutions between the clustered and original models were. Additionally, the consideration of uncertainty in the energy hub modeling was proven to be beneficial, particularly when the environmental constraints were included in the formulation. In addition, the heuristic approach remarkably outperformed the full-scale model in terms of the solution time, usually by several orders of magnitude, particularly for large datasets. Accordingly, it can be stated that employing the clustering approach is an effective tool to reduce the model size while maintaining reasonable, accurate results. The proposed multiscale clustering method is a trade-off between the computational effort and data accuracy.

Future works can include the application of the proposed clustering approach to different multiscale planning problems. The stochastic energy hub planning model can be extended to include capacity expansion planning decisions to satisfy the multiple attributes demand. It would be interesting to use forecasted demand data to plan energy hub systems, as this case study was limited to implementing historical demand data into multiscale modeling. Therefore, forecasting techniques can be employed to forecast the future demands; the clustering approach will be applied to reduce the size of these multiple attributes demands where they can be used as an input to the energy hub planning or capacity expansion model. Another example for future work is that the multiscale clustering approach can be applied to a superstructure modeling approach to design new chemical or power plants. Therefore, instead of solving the superstructure model for a 1-day profile that represents the whole year, it can be solved for several representative days that are more likely to reflect the real behavior of demand.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/pr11041046/s1>.

Author Contributions: Conceptualization, M.A., A.B.-T. and F.A.; methodology, M.A.; software M.A. validation, M.A. and A.E.; formal analysis, M.A.; investigation, M.A.; resources, M.A. and F.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A. and A.B.-T.; visualization, M.A.; supervision, A.E. and A.A.; project administration, A.E. and A.A.; funding acquisition, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weinan, E. *Principles of Multiscale Modeling*; Cambridge University Press: Cambridge, UK, 2011.
2. Yilmaz, H.; Fouché, E.; Dengiz, T.; Krauß, L.; Keles, D.; Fichtner, W. Reducing energy time series for energy system models via self-organizing maps. *Inf. Technol.* **2019**, *61*, 125–133. [[CrossRef](#)]
3. Moghaddam, I.G.; Saniei, M.; Mashhour, E. A comprehensive model for self-scheduling an energy hub to supply cooling, heating and electrical demands of a building. *Energy* **2016**, *94*, 157–170. [[CrossRef](#)]
4. Ma, T.; Wu, J.; Hao, L. Energy flow modeling and optimal operation analysis of the micro energy grid based on energy hub. *Energy Convers. Manag.* **2017**, *133*, 292–306. [[CrossRef](#)]
5. Lu, Q.; Lü, S.; Leng, Y.; Zhang, Z. Optimal household energy management based on smart residential energy hub considering uncertain behaviors. *Energy* **2020**, *195*, 117052. [[CrossRef](#)]
6. Taqvi, S.T.; Almansoori, A.; Maroufmashat, A.; Elkamel, A. Utilizing Rooftop Renewable Energy Potential for Electric Vehicle Charging Infrastructure Using Multi-Energy Hub Approach. *Energies* **2022**, *15*, 9572. [[CrossRef](#)]
7. Guo, W.; Wang, Q.; Liu, H.; Desire, W.A. Multi-energy collaborative optimization of park integrated energy system considering carbon emission and demand response. *Energy Rep.* **2023**, *9*, 3683–3694. [[CrossRef](#)]
8. Koltsaklis, N.E.; Kopanos, G.M.; Georgiadis, M.C. Design and Operational Planning of Energy Networks Based on Combined Heat and Power Units. *Ind. Eng. Chem. Res.* **2014**, *53*, 16905–16923. [[CrossRef](#)]
9. Maroufmashat, A.; Sattari, S.; Roshandel, R.; Fowler, M.; Elkamel, A. Multi-objective Optimization for Design and Operation of Distributed Energy Systems through the Multi-energy Hub Network Approach. *Ind. Eng. Chem. Res.* **2016**, *55*, 8950–8966. [[CrossRef](#)]
10. Mukherjee, U.; Maroufmashat, A.; Narayan, A.; Elkamel, A.; Fowler, M. A Stochastic Programming Approach for the Planning and Operation of a Power to Gas Energy Hub with Multiple Energy Recovery Pathways. *Energies* **2017**, *10*, 868. [[CrossRef](#)]
11. Kotzur, L.; Markewitz, P.; Robinius, M.; Stolten, D. Impact of different time series aggregation methods on optimal energy system design. *Renew. Energy* **2018**, *117*, 474–487. [[CrossRef](#)]
12. Zhang, H.; Tomasgard, A.; Knudsen, B.R.; Svendsen, H.G.; Bakker, S.J.; Grossmann, I.E. Modelling and analysis of offshore energy hubs. *Energy* **2022**, *261*, 125219. [[CrossRef](#)]
13. Amry, Y.; Elbouchikhi, E.; Le Gall, F.; Ghogho, M.; El Hani, S. Optimal sizing and energy management strategy for EV workplace charging station considering PV and flywheel energy storage system. *J. Energy Storage* **2023**, *62*, 106937. [[CrossRef](#)]
14. Zhang, T.; Wang, M.; Wang, P.; Gu, J.; Zheng, W.; Dong, Y. Bi-stage stochastic model for optimal capacity and electric cooling ratio of CCHPs—A case study for a hotel. *Energy Build.* **2019**, *194*, 113–122. [[CrossRef](#)]
15. Faraji, J.; Hashemi-Dezaki, H.; Ketabi, A. Stochastic operation and scheduling of energy hub considering renewable energy sources' uncertainty and N-1 contingency. *Sustain. Cities Soc.* **2020**, *65*, 102578. [[CrossRef](#)]
16. Majidi, M.; Nojavan, S.; Zare, K. A cost-emission framework for hub energy system under demand response program. *Energy* **2017**, *134*, 157–166. [[CrossRef](#)]
17. Nojavan, S.; Majidi, M.; Zare, K. Optimal scheduling of heating and power hubs under economic and environment issues in the presence of peak load management. *Energy Convers. Manag.* **2018**, *156*, 34–44. [[CrossRef](#)]
18. Maroufmashat, A.; Fowler, M.; Khavas, S.S.; Elkamel, A.; Roshandel, R.; Hajimiragha, A. Mixed integer linear programming based approach for optimal planning and operation of a smart urban energy network to support the hydrogen economy. *Int. J. Hydrogen Energy* **2016**, *41*, 7700–7716. [[CrossRef](#)]
19. Wang, H.; Zhang, H.; Gu, C.; Li, F. Optimal design and operation of CHPs and energy hub with multi objectives for a local energy system. *Energy Procedia* **2017**, *142*, 1615–1621. [[CrossRef](#)]
20. Vahid-Pakdel, M.; Nojavan, S.; Mohammadi-Ivatloo, B.; Zare, K. Stochastic optimization of energy hub operation with consideration of thermal energy market and demand response. *Energy Convers. Manag.* **2017**, *145*, 117–128. [[CrossRef](#)]
21. Alhameli, F.; Elkamel, A.; Betancourt-Torcat, A.; Almansoori, A. A mixed-integer programming approach for clustering demand data for multiscale mathematical programming applications. *AIChE J.* **2019**, *65*, e16578. [[CrossRef](#)]
22. Turk, A.; Wu, Q.; Zhang, M.; Østergaard, J. Day-ahead stochastic scheduling of integrated multi-energy system for flexibility synergy and uncertainty balancing. *Energy* **2020**, *196*, 117130. [[CrossRef](#)]
23. Elkamel, M.; Valencia, A.; Zhang, W.; Zheng, Q.P.; Chang, N.-B. Multi-agent modeling for linking a green transportation system with an urban agriculture network in a food-energy-water nexus. *Sustain. Cities Soc.* **2023**, *89*, 104354. [[CrossRef](#)]
24. Liu, T.; Zhang, D.; Wang, S.; Wu, T. Standardized modelling and economic optimization of multi-carrier energy systems considering energy storage and demand response. *Energy Convers. Manag.* **2019**, *182*, 126–142. [[CrossRef](#)]
25. Bektaş, S.; Şişman, Y. The comparison of L11 and L22-norm minimization methods. *Int. J. Phys.* **2010**, *5*, 1721–1727.
26. Chelmiss, C.; Kolte, J.; Prasanna, V.K. Big data analytics for demand response: Clustering over space and time. In Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015. [[CrossRef](#)]
27. Green, R.; Staffell, I.; Vasilakos, N. Divide and Conquer? k -Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System. *IEEE Trans. Eng. Manag.* **2014**, *61*, 251–260. [[CrossRef](#)]
28. Lyu, Q.; Lin, Z.; She, Y.; Zhang, C. A comparison of typical ℓ minimization algorithms. *Neurocomputing* **2013**, *119*, 413–424. [[CrossRef](#)]
29. Sabo, K. Center-based l_1 -clustering method. *Int. J. Appl. Math. Comput. Sci.* **2014**, *24*, 151–163. [[CrossRef](#)]

30. Branke, J.; Deb, K.; Miettinen, K.; Słowiński, R. (Eds.) *Multiobjective Optimization: Interactive and Evolutionary Approaches*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5252.
31. Vinod, H.D. Integer Programming and the Theory of Grouping. *J. Am. Stat. Assoc.* **1969**, *64*, 506. [[CrossRef](#)]
32. Mangasarian, O.L. Absolute value equation solution via dual complementarity. *Optim. Lett.* **2013**, *7*, 625–630. [[CrossRef](#)]
33. Mirzaesmaeeli, H.; Elkamel, A.; Douglas, P.; Croiset, E.; Gupta, M. A multi-period optimization model for energy planning with CO₂ emission consideration. *J. Environ. Manag.* **2010**, *91*, 1063–1070. [[CrossRef](#)]
34. Sağlam, B.; Salman, F.S.; Sayın, S.; Türkay, M. A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *Eur. J. Oper. Res.* **2006**, *173*, 866–879. [[CrossRef](#)]
35. Üney, F.; Türkay, M. A mixed-integer programming approach to multi-class data classification problem. *Eur. J. Oper. Res.* **2006**, *173*, 910–920. [[CrossRef](#)]
36. GAMS Development Corporation. *General Algebraic Modeling System (GAMS) Release 23.3.3*; GAMS: Washington, DC, USA; Cologne, Germany, 2009.
37. da Rosa, A. Wind Energy. In *Fundamentals of Renewable Energy Processes*; Elsevier: Amsterdam, The Netherlands, 2013; pp. 685–763. [[CrossRef](#)]
38. Sengupta, M.; Xie, Y.; Lopez, A.; Habte, A.; Maclaurin, G.; Shelby, J. The National Solar Radiation Data Base (NSRDB). *Renew. Sustain. Energy Rev.* **2018**, *89*, 51–60. [[CrossRef](#)]
39. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
40. Maroufmashat, A.; Elkamel, A.; Fowler, M.; Sattari, S.; Roshandel, R.; Hajimiragha, A.; Walker, S.; Entchev, E. Modeling and optimization of a network of energy hubs to improve economic and emission considerations. *Energy* **2015**, *93*, 2546–2558. [[CrossRef](#)]
41. Stander, J. The Specification of a Small Commercial Wind Energy Conversion System for the South African Antarctic Research Base SANAE IV. Master's Thesis, Stellenbosch University, Stellenbosch, South Africa, 2008.
42. Battelle Memorial Institute. *Manufacturing Cost Analysis: 100kW and 250 kW Fuel Cell Systems for Primary Power and Combined Heat and Power Applications*; U.S. Department of Energy, Fuel Cell Technologies Office: Danbury, CT, USA, 2017.
43. Birge, J.R.; Louveaux, F. *Introduction to Stochastic Programming*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2011.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.